Jeremy Holleman
Fan Zhang
Brian Otis

Ultra Low-Power Integrated Circuit Design for Wireless Neural Interfaces

Springer

Ultra Low-Power Integrated Circuit Design for Wireless Neural Interfaces

Jeremy Holleman • Fan Zhang • Brian Otis

Ultra Low-Power Integrated Circuit Design for Wireless Neural Interfaces

Springer

Jeremy Holleman
Department of Electrical Engineering
and Computer Science
University of Tennessee
414 Ferris Hall
Knoxville, TN 37996-2100, USA
hollemj@u.washington.edu

Brian Otis
Department of Electrical Engineering
University of Washington
185 Stevens Way
Seattle, WA 98195-2500, USA
botis@u.washington.edu

Fan Zhang
Department of Electrical Engineering
University of Washington
185 Stevens Way
Seattle, WA 98195-2500, USA
fz2@u.washington.edu

ISBN 978-1-4419-6726-8 e-ISBN 978-1-4419-6727-5
DOI 10.1007/978-1-4419-6727-5
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2010936899

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Contents

Chapter 1
Introduction

Neuroscientists are increasingly engaging the integrated circuit (IC) community to develop new tools for understanding the brain. Fundamental research performed on small animal models, for example, requires miniaturized instrumentation for long term freely behaving studies. Recording from non-human primates, rats, mice, and even insects is of interest. This research, in turn, will lead to advanced neuroprosthetics and brain-computer interfaces (BCI), which will demand even more functionality, robustness, and miniaturization from the electronics. Overly conservative performance goals lead to a loss of efficiency, while overly relaxed specifications lead to an ineffective system. Since there are no established standards, close interaction between IC designers and neuroscientists is critical. Our goal with this book is to present several case studies of low power circuit architectures that were designed for brain interface applications. Our chip specifications, design procedures, and measured results will be presented. Along the way, we hope to motivate the need for additional research and collaborations between engineers and brain scientists.

These collaborations have already led to important fundamental scientific research. For example, neural interfaces have facilitated discoveries in basic neuroscience research [4] by enabling previously impossible experiments. In the short term, these technologies are instrumental in developing implantable wireless sensors with a small form-factor and low weight. Wireless sensors will facilitate advanced biomedical research, like untethered monitoring of freely-behaving insects and small animals (Fig. 1.1a) [5]. These systems require wireless information transfer between implanted electrodes and external devices. In the long term, brain-machine interfaces (BMI) may provide new augmentative technology to help people with severe motor disabilities (Fig. 1.1b) [10]. BMIs acquire brain signals and extract specific features from them, then translate them into device control signals (e.g., controlling a mouse cursor). To people who lack any useful muscle control (e.g., due to cerebral palsy), or locked-in (e.g., end-stage ALS), BMIs could give the ability to answer simple questions quickly, control the environment, or even operate neuroprosthetic devices.

Prototype systems have demonstrated the potential to profoundly improve the quality of life for persons with severe impairments [13]. Medical applications have begun to appear, starting with the cochlear implant over three decades ago [8].

J. Holleman et al., *Ultra Low-Power Integrated Circuit Design for Wireless Neural Interfaces*,
DOI 10.1007/978-1-4419-6727-5_1, © Springer Science+Business Media, LLC 2011

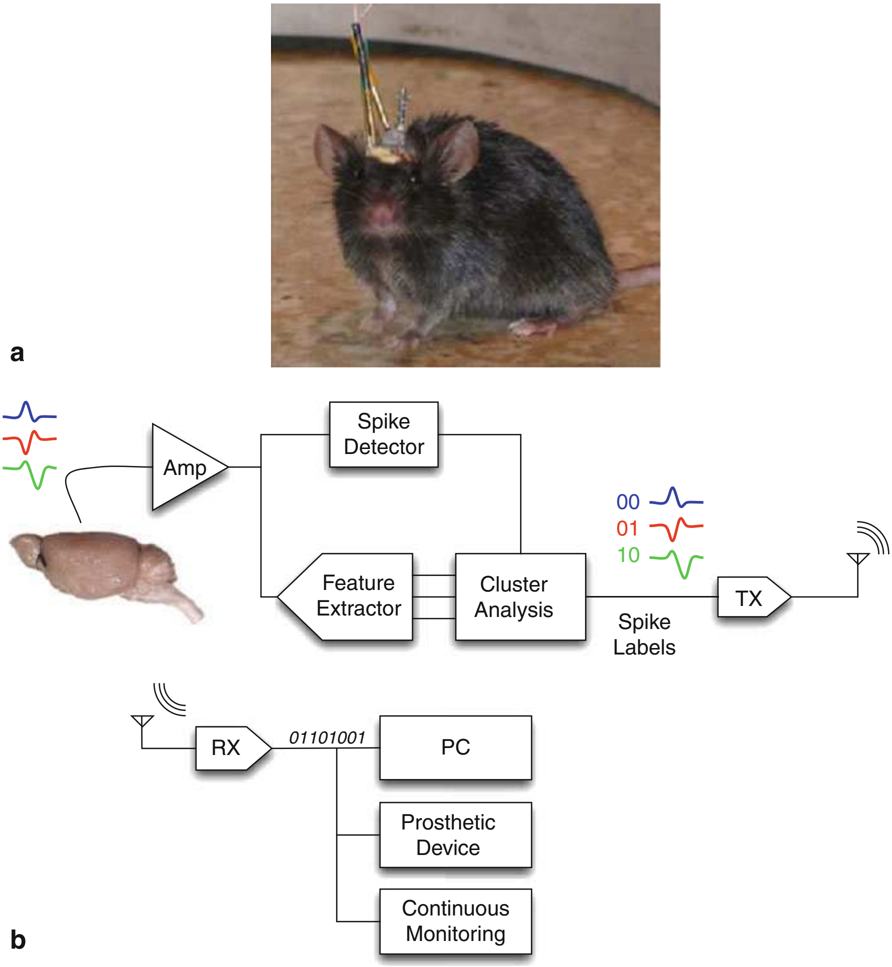


Fig. 1.1 **a** Short-term biomedical applications of microelectronics can enable advances in biology and medical research. For example, the wired connection to study mice shown here can be replaced with a wireless device, allowing the study of freely-behaving mice. **b** A typical neural recording architecture illustrating the concept of Brain-Machine-Interfaces (BMI) and Long-term biomedical applications of microelectronics

Currently, work is ongoing for neural interfaces to improve treatment for epilepsy and Parkinson's disease. Scientists and engineers are also investigating the possibility of neurally-controlled prosthetics: devices which would allow persons paralyzed by injury or disease to control a prosthetic through thoughts. Rudimentary neural control has been demonstrated both in humans [7] and in monkeys [12].

Current neural interfaces are limited by physical size and power consumption. One challenge in reducing power consumption is the noise requirement of the first gain stage. The small amplitude of extracellular-sensed neural signals ($<200\,\mu\text{V}$)

necessitates low-noise amplification, which in turn requires relatively high bias currents. While some components of a neural interface, such as a wireless transceiver, may operate with a low duty cycle, the amplifiers must operate continuously. The combination of high duty cycle and low noise requirements frequently cause the amplifiers to dominate overall system power, particularly in systems with many channels.

An additional obstacle to reducing power consumption is the need for local, real-time processing of neural signals [6]. One particularly difficult and important analysis function for neural recording is spike sorting. Electrodes implanted in a brain can often detect action potentials from multiple neurons. Spike sorting, the task of distinguishing between the different neurons contributing to activity in a recording, can increase the information that a recording provides compared to simple thresholding algorithms. This extra information can improve the performance of medical devices such as neurally-controlled prosthetics, and improve the ability of neuroscientists to infer the meaning of experiments.

Unfortunately, spike sorting is a difficult function to implement within the power and area constraints of an implanted neural interface. Many spike sorting algorithms require extensive memory to store ensembles of recorded spikes, which consumes large amounts of chip area and power. For these reasons, a fully-integrated neural recording IC with on-chip spike-sorting has not yet been developed.

Figure 1.2 shows three possible architectures for an implantable neural recording system. In all cases, the first stage is a low-noise amplifier. One option, shown in Fig. 1.2a is to fully digitize the signals and transmit them to an external computer for processing. This strategy benefits from the flexibility and processing power of general-purpose computers. Users can choose processing algorithms to suit their needs and modify the algorithms at any time. Additionally, since no on-line spike sorting is performed, the raw data is available and any questions about the accuracy of the processing can be resolved. This architecture requires that analog-digital converters (ADC) run continuously for every channel. The illustration shows an independent ADC for every channel. A typical implementation would have several channels multiplexed to share an ADC, but a large number of channels would still require multiple area-intensive ADCs. While channel-multiplexing can reduce the area required for ADCs, it increases power consumption due to the additional buffering required [2], so that the power required for conversion can still be expected to scale at least linearly with the number of channels. The power required for analog-digital conversion can consume a significant portion of the system's power budget. Even more problematic is the wireless transmission. Because the full digitized waveform for every channel is transmitted over the wireless link, the transmitter must operate continuously and must also have very high throughput, resulting in prohibitively high power dissipation. For example, two recently published transmitters operating in the Medical Implant Communication Service (MICS) band, [1] and [9], achieved energy-per-transmitted-bit of 2.9 and 4 nJ/b and data rates of 120 and 100 kb/s, respectively. For a 100-channel system using 8-bit digitization and a 30 kS/s sample rate, similar transmitter efficiency would require 70–95 mW for the transmitter alone. For reference, an implant with 1 cm^2 of surface area can dissipate no more than 80 mW without risking thermal damage to tissue [11].

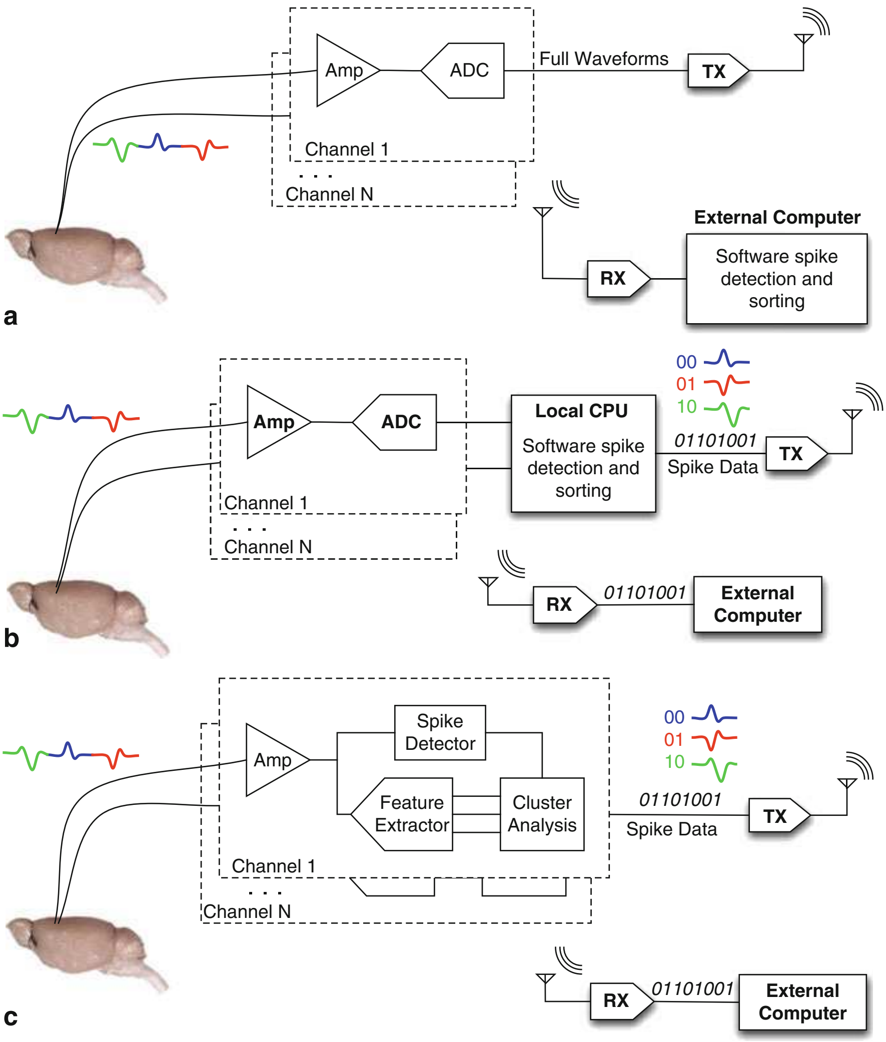


Fig. 1.2 Possible architectures for a wireless neural interface. **a** The signals are digitized and a wireless transmitter (Tx) sends the waveforms for all channels to an external computer, where they are processed in software. **b** All of the signals are digitized, then processed in software on a local CPU which is part of the implant. The output of the spike processing, which includes spike timestamps, sorting labels, and a channel index, are transmitted to an external computer, where they are collected for analysis or used to actuate a prosthetic. **c** Each channel is processed locally using dedicated analog circuits. As in (**b**), the processing results are transmitted to an external computer

A second possibility, depicted in Fig. 1.2b, is to digitize the signals and process them locally. A local CPU or DSP would detect spikes, perform spike sorting, and record the time and channel for each detected spike. The resulting spike data would then be transmitted to an external computer, where it could be collected for further analysis in the context of neuroscience research, or used to actuate a prosthesis. Such a strategy requires the same power and area for digitization as the previous architecture. The demands on the wireless transmitter are dramatically reduced since each full waveform is replaced with a set of spike descriptors. In exchange for reduced transmission, this architecture requires substantial local processing power. The CPU included in the implant must be capable of performing spike detection and sorting on all channels simultaneously and in real time. The chip must also contain sufficient memory to hold the processing software.

The processing can also be done locally using dedicated analog circuits, as shown in Fig. 1.2c. Similarly to the design in Fig. 1.2b, the burden on the communication link is relatively light because only spike descriptors are transmitted. Unlike either of the other two architectures, there is no need for a continuously-running ADC, since spikes are detected and sorted in the analog domain. In fact, the ADC could be omitted entirely, although it may be desirable to include one which can be enabled periodically to compare processing results with the raw waveform.

As noted in the above discussion, any architecture for an implantable neural interface will require a low-noise amplifier for every active channel. One of the goals of this book is the development of circuit techniques to reduce power dissipation in such an amplifier. The other primary contribution is the design of circuits which exploit the natural behavior of transistors to perform analog computation implementing spike detection, feature extraction, and clustering. Among these circuits is the first reported floating-gate memory cell using thin-oxide transistors for adaptation, which enables floating-gate circuits for machine learning algorithms to operate from supply voltages of 1.5 V. These cells are used to store spike templates, enabling a fully analog spike sorting circuit.

These circuits are the critical building blocks for an implantable neural interface using the architecture shown in Fig. 1.2c. Because they operate with extremely low power dissipation, they will enable improved implantable devices which can be used in many demanding applications.

To emphasize the circuit design challenges presented by this vision, we will briefly describe some of the specifications of the front-end neural recording amplifiers. Like the other building blocks described above, several design procedures and case studies of neural recording amplifiers will be presented in this book.

Bio-signals need to be first amplified before digitization or any signal processing. Depending on the application, several design requirements should be satisfied for the front-end amplifiers:

- Have sufficiently low input-referred noise to resolve microvolt-level spikes (10 μV)
- Have sufficient dynamic range to convey or tolerate large local field potential (LFP) or EMG (muscle) signals (1–10 mV)
- Have much higher input impedance than the electrode-tissue interface and negligible DC input current;
- Amplify signals in the frequency band of interest (300–5 kHz for spikes, 10–100 Hz for LFP, 0.5–40 Hz for EEG, 0.5–200 Hz for ECoG, and 0.5–20 Hz for EMG).

- Block (or cancel) DC offsets present at the electrode-tissue interface to prevent saturation at the amplifier output.
- Consume as little silicon area as possible and use few or no off-chip components to minimize size.
- Sufficient common-mode rejection ratio (CMRR) to minimize interference from 50/60 Hz power line noise, and sufficient power supply rejection ratio (PSRR) to prevent coupling from power supply noise (more severe if inductive power link is used).
- Dissipate no more than 10 mW of power.

The critical power limitation on the implantable recording electronics arises from the need to limit the chronic heating of surrounding tissue to less than 1°C. Preliminary experiments have shown that an implanted cortical 100-electrode array with integrated electronics can safely dissipate approximately 10 mW of power [3]. It follows that each channel must consume less than 100 µW of power, excluding shared circuit blocks such as analog-to-digital converter (ADC), power regulation and transmitter.

The first few chapters describe techniques for designing neural amplifiers with low power consumption and low noise. Chapter 2 begins by describing the challenges involved in neural amplifier design. Chapter 3 describes a simple open-loop topology that achieves excellent noise/power performance, in part by sacrificing other metrics. Chapters 4 and 5 describe the design and measurements of two amplifiers. One combines a traditional architecture with low-voltage design techniques. The other incorporates the insights from the open-loop design and the low-voltage design just mentioned into a low-voltage closed-loop amplifier with very low noise and good overall performance.

The next three chapters describe circuit techniques for processing neural signals. Chapter 7 presents a chip which detects neural spikes, extracts descriptive features, and digitizes the features. Chapter 8 describes the process of spike sorting and some considerations for an analog implementation of a spike-sorting algorithm. Chapter 9 describes a circuit that uses thin-oxide floating-gate analog memories to realize an unsupervised clustering algorithm, which is the primary component of a spike sorting system.

Finally, the last two chapters examine system-level integration in the context of two example systems. One is the NeuralWISP, a wirelessly-powered spike density recorder designed to work with a commercial RFID reader. The other is a neural streaming chip, which amplifies, digitizes and wirelessly transmits a neural signal. These systems are described in Chaps. 10 and 11 respectively.

References

[1] Bohorquez J, Dawson J, Chandrakasan A (2008) A 350 µW CMOS MSK transmitter and 400 µW OOK super-regenerative receiver for medical implant communications. In: IEEE symposium on VLSI circuits, Digest of Technical Papers, Massachusetts Institute of Technology, Cambridge, MA, pp 32–33

[2] Chae M, Liu W, Sivaprakasam M (2008) Design optimization for integrated neural recording systems. IEEE J Solid-State Circuits 43(9):1931–1939

[3] Harrison R (2008) The design of integrated circuits to observe brain activity. Proc IEEE 96(7):1203–1216

[4] Jackson A, Mavoori J, Fetz E (2006) Long-term motor cortex plasticity induced by an electronic neural implant. Nature 444:56–60

[5] Kipke D, Shain W, Buzsaki G, Fetz E, Menderson J, Hetke J, Schalk G (2008) Advanced neurotechnologies for chronic neural interfaces: new horizons and clinical opportunities. J Neurosci 28(46):11830–11838

[6] Lebedev M, Nicolelis M (2006) Brain–machine interfaces: past, present and future. Trends Neurosci 29(9):536–546

[7] Leuthardt E, Miller K, Schalk G, Rao R, Ojemann J (2006) Electrocorticography-based brain computer interface—the seattle experience. IEEE Trans Neural Syst Rehabil Eng 14(2): 194–198

[8] Michelson R, Merzenich M, Pettit C, Schindler R (1973) A cochlear prosthesis: further clinical observations; preliminary results of physiological studies. Laryngoscope 83(7):1116–1122

[9] Rai S, Holleman J, Pandey J, Zhang F, Otis B (2009) A 500 μW neural tag with 2 μV rms AFE and frequency multiplying MICS/ISM FSK transmitter. In: IEEE international solid-state circuits conference, Digest of Technical Papers, pp 212–213

[10] Schalk G, McFarland D, Hinterberger T, Birbaumer N, Wolpaw J (2004) Bci2000: a general-purpose brain-computer interface (bci) system. IEEE Trans Biomed Eng 51(6):1034–1043 URL: http://www.kl-ic.com/white9.pdf

[11] Seese T, Harasaki H, Saidel G, Davies C (1998) Characterization of tissue morphology, angiogenesis, and temperature in the adaptive response of muscle tissue to chronic heating. Laboratory investigation; J Tech Methods Pathol 78(12):1553–1562

[12] Wessberg J, Stambaugh C, Kralik J, Beck P, Laubach M, Chapin J, Kim J, Biggs S, Srinivasan M, Nicolelis M (2000) Real-time prediction of hand trajectory by ensembles of cortical neurons in primates. Nature 408:361–365

[13] Wise K, Anderson D, Hetke J, Kipke D, Najafi K (2004) Wireless implantable microsystems: high-density electronic interfaces to the nervous system. Proc IEEE 92(1):76–97

Chapter 2
Bio-Signal Interface Amplifiers: An Introduction

There are many design challenges involved in the circuit design of implantable neural recording systems. A generic biopotential-recording system is illustrated in Fig. 2.1. First, weak neural signals must be amplified, conditioned, and then digitized. The information then needs to be wirelessly transmitted out of the body to avoid possible infection from transcutaneous connectors. The power consumption increases with the number of recording channels and the complexity of system. However, the power dissipation of miniature implantable devices must be limited to prevent excessive tissue heating.

In a typical multi-channel system, one distinct low-noise amplifier is used per signal channel. Analog multiplexing theoretically would reduce the number of front-end amplifiers. However, in order to capture details of ever-changing neural activity across multiple electrodes, analog multiplexing requires switching times much shorter than the time constants associated with the amplifier's dynamics. Therefore, multi-channel systems typically use a separate amplifier for each channel, severely limiting the power available for each amplifier. As a result, power dissipation must be minimized as much as possible.

The next few chapters of this book will go into detail on ultra-low power low noise amplifier (LNA) design used in neural recording and other bio-signal acquisition systems. We first begin with the signal and electrode characteristics of these systems.

2.1 Characteristics of the Recording Electrodes

A signal/reference electrode configuration is typically used to record neural or muscle activity. The potential difference between each signal electrode and a large reference electrode is measured by the front-end differential amplifiers. The reference electrodes are usually low-impedance. However, in other applications, some signal electrodes are paired with high-impedance reference electrodes. The contact between metal electrode tip and extracellular fluid creates an electrical double layer that acts like a capacitor. Depending on the electrode area and surface roughness, the capacitance is estimated between 150 pF and 1.5 nF in common electrodes.

J. Holleman et al., *Ultra Low-Power Integrated Circuit Design for Wireless Neural Interfaces*,
DOI 10.1007/978-1-4419-6727-5_2, © Springer Science+Business Media, LLC 2011

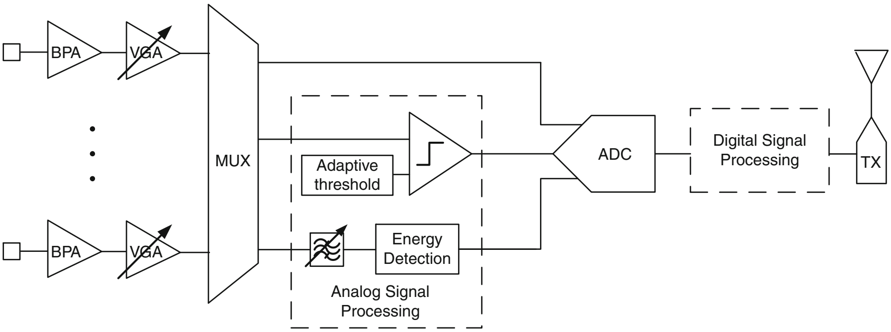


Fig. 2.1 A generic block diagram for a biopotential-recording system

Recent advances in MEMS technology have produced small (less than 4 mm in each dimension) arrays of micro-electrodes containing as many as 100 recording sites [11].

Below we illustrate electrodes commonly used in research laboratories. Figure 2.2 shows the neural electrodes from NeuroNexus Technology. There are a total of 128

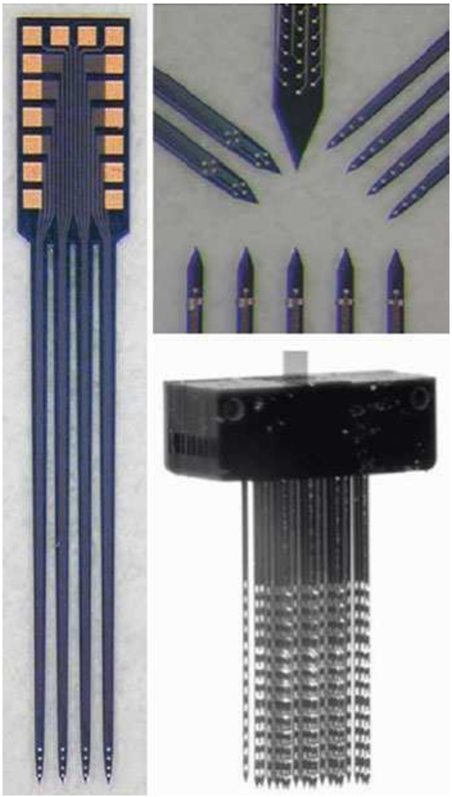


Fig. 2.2 Typical invasive neural electrodes often used for in-vivo recording on rats or monkeys

sites with 200 μm electrode spacing. The electrodes used to acquire the measurements in this book and in [6] are 50 μm tungsten wires insulated with teflon. Each electrode is capacitive and has an equivalent impedance magnitude of 100 k–500 kΩ measured at 1 kHz.

2.2 Characteristics of Bio-Signals

2.2.1 Brain Recordings

There are three main types of invasive signals that are of interest: action potentials (spikes), local field potentials (LFP), and electrocorticography (ECoG) signals. Spikes and LFPs can be obtained from single-unit recording. LFPs can be measured on the scalp as EEG signals, but experience a significant amount of attenuation. ECoG signals can be measured by invasive recording electrodes at the surface of the cortex. Typical spikes have signals occupying the 100 Hz–7 kHz band with amplitudes up to 500 μV, while LFPs generally have energy below 100 Hz with amplitudes up to 5 mV [10]. Related to LFPs, EEG recordings have signals much attenuated to 10–20 μV. ECoG signals have energy in roughly the 0.5–200 Hz band with amplitudes up to 100 μV.

Neural spikes appear biphasic in in-vivo recordings with durations of 0.3–1.0 ms. The spikes fire once every several milliseconds to tens of milliseconds depending on the location of the electrode and the neuron's inherent characteristics. Spikes from different neurons usually have different shapes and firing rates, whereas spikes from the same neuron have nearly identical amplitude and duration. Spikes provide high spatial resolution at the cost of high power consumption in the recording electronics and challenging chronic implanting issues at the electrode-tissue interface.

LFPs result from the collective activity of many neurons in one region of the brain. Some neurons are too distant from the electrode to have their individual spikes resolved. LFPs have much less spatial resolution compared to neural spikes; however, it is more immune to attenuation (i.e., caused by scar tissue) and interference. Some research has demonstrated close correlation between specific arm movement and the energy of LFP signals in primates [3].

Another type of signal that is recently gaining popularity is categorized as electrocorticography (ECoG). ECoG signals are generally recorded from the surface of the cortex, and are thus less susceptible to chronic measurement issues such as tissue encapsulation and micromotion [2]. As a result, they can provide more robust measurement of signals. Although less spatially refined than spikes recorded from single-cell recordings, they are more spatially refined comparing to EEG signals (tenths of millimeters vs. centimeters [7]. Some recent research has demonstrated the effective usage of spectral decomposition of ECoG signals in neuroprosthetic applications [9, 13]. Ensemble neural firing is the biomarker for a number of clinically relevant phenomena, such as epileptic seizures and basal ganglia rhythms in Parkinson's disease [8].

Table 2.1 Characteristics of biosignals

	Bandwidth	Amplitude	Spatial Resolution	Invasiveness
Spikes	1–7 kHz	<500 μV	Highest	Invasive
LFP	<200 Hz	<5 mV	Low	Invasive
EEG	<100 Hz	10–20 μV	Lowest	Non-invasive
ECoG	0.5–200 Hz	<100 μV	Moderate	Moderately invasive
EMG	7–500 Hz	50 μ–2 mV	–	Minimally or non-invasive

2.2.2 *Muscle-Based Signals*

Electromyography (EMG) is another kind of signal measured from muscle cells. Measured EMG potentials range between 50 μV up to 30 mV in the band of 7–500 Hz, depending on the particular muscle. EMG signals are measured either by surface recording or needle (intramuscular) EMG. EMG signals are used in many types of research laboratories, including bio-mechanics, motor control, neuromuscular physiology, etc. Clinically, they are also used for the diagnosis of neurological and neuromuscular problems.

Table 2.1 is a summary of the characteristics of biosignals.

2.3 Noise/Power Tradeoff

As explained previously, reducing the power consumption of the circuitry is imperative to allow a practical multichannel implantable system. In order to understand how to optimize the power and noise trade-off, we first investigate types of noise and their relationship to bias current.

There are mainly two noise sources that circuit designers consider when designing low-frequency low-noise amplifiers: flicker noise and thermal noise.

2.3.1 *Flicker Noise, 1/f Noise*

Flicker noise is thought to be caused by traps associated with contamination and crystal defects [4]. These traps capture and release carriers randomly and give rise to a noise signal with energy concentrated at low frequencies. Input-referred flicker noise of a MOS can be represented by

$$\overline{v_i^2} = \left(\frac{K_f}{WLC_{ox} f}\right) \Delta f \tag{2.1}$$

2.3.2 Thermal Noise

In conventional resistors, thermal noise is due to the random thermal motion of the electrons, thus is directly proportional to T. Input-referred thermal noise voltage noise of a MOSFET can be represented by

$$\overline{v_i^2} = 4kT\left(\frac{2}{3g_m}\right)\Delta f \tag{2.2}$$

A common dimensionless figure of merit that summarizes this power-noise trade-off is the noise efficiency factor (NEF), first proposed in [14].

$$\mathrm{NEF} = V_{rms,in}\sqrt{\frac{2\cdot I_{Total}}{\pi\cdot U_T\cdot 4kT\cdot BW}} \tag{2.3}$$

Where I_{Total} is the total amplifier supply current, U_T is the thermal voltage k_T/q, BW is the amplifier bandwidth, and $V_{ni,rms}$ is the amplifier's input-referred RMS voltage noise. This FOM compares the power-noise trade-off with that of a single ideal bipolar transistor.

2.4 Representative Prior Art

Micro-scale integrated circuits used for amplifying weak bioelectrical signals have been reported for many years [10, 11, 14]. Since then, many papers on low-noise amplifiers have been published [1, 10, 12, 14]. In 2003, [5] reported a fully integrated amplifier consuming 80 µW of power while exhibiting an RMS input-referred noise of 2.2 µV. Their impressive noise efficiency factor (4.0) has set the benchmark for amplifier designers of that time.

The amplifier in [5] is based on a simple operational transconductance amplifier (OTA) topology. A capacitive feedback network sets the midband gain of the amplifier. Any DC offset from the electrode-tissue interface is removed by capacitively coupling the inputs through capacitors. Large pseudo-resistors are used to set the low-frequency amplifier cutoff. The pseudo-resistors are used in place of bulky resistors in order to save area. They are MOS-bipolar elements that create a small-signal resistance of $>10^{12}$ Ω for low-frequency operation. The noise optimization is accomplished by sizing the input stage transistors to operate in the sub-threshold region.

The topology in [5] is not optimal for a power-noise trade-off as 50% of the current is consumed in a branch that contributes no gain. The design in [15] employs a modified folded-cascode topology where the currents between the input and the folded branches are severely scaled (17:1) to save power and reduce the noise contribution from the folded branches. In addition, source degeneration is used at the input current sources to increase their output impedance and reduce their noise contribution. However, a folded-cascode topology is not a power-efficient solution to achieve low noise because its extra branches consume more current and contribute more noise. In addition, the amplifier in [5] uses a ±5 V, and [15] uses 2.8 V supply. With the trend towards integrating analog and digital subsystems on a single die, it has become increasingly important for analog circuitry to operate from the <1.2 V supply typical of modern digital CMOS.

In the next few chapters of this book, we will introduce several new amplifier topologies and compare their measured results to the state-of-the-art.

References

[1] Chandran A, Najafi K, Wise K (1999) A new DC baseline stabilization scheme for neural recording microprobes. In: Proceedings of the first joint BMES/EMBS conference, annual fall meeting of the Biomedical Engineering Society, 21st annual international conference of the Engineering in Medicine and Biology Society, Atlanta, Georgia, pp 386–387

[2] Denison T, Consoer K, Santa W, Avestruz A, Cooley J, Kelly A (2009) A 2 μW, 100 nV/rtHz chopper-stabilized instrumentation amplifier for chronic measurement of neural field potentials. IEEE J Solid-State Circuits 42(12):2934–2945

[3] Donoghue J, Sanes J, Hatsopoulos N, Gaal G (1998) Neural discharge and local field potential oscillations in primate motor cortex during voluntary movements. J Neurophysiol 79:159–173

[4] Gray P, Hurst PJ, Lewis S, Meyer R (2009) Analysis and design of analog integrated circuits. Wiley, Berkeley

[5] Harison R, Charles C (2003) A low-power low-noise cmos amplifier for neural recording applications. IEEE J Solid-State Circuits 39:122–131

[6] Kipke D, ShainW, Buzsaki G, Fetz E, Menderson J, Hetke J, Schalk G (2008) Advanced neurotechnologies for chronic neural interfaces: new horizons and clinical opportunities. J Neurosci 28(46):11830–11838

[7] Leuthardt E, Schalk G, Wolpaw J, Ojemann J, Moran D (2004) A braincomputer interface using electrocorticographic signals in humans. J Neural Eng 1(2):63–71

[8] Levy R, Ashby P, Hutchison W, Lang A, Lozano A, Dostrovsky J (2002) Dependence of subthalamic nucleus oscillations on movement and dopamine in parkinson's disease. Brain 125:1175–1176

[9] Miller K, Leuthardt E, Schalk G, Rao R, Anderson N, Moran D, Miller J, Ojemann J (2009) Spectral changes in cortical surface potentials during motor movement. J Neurosci 27(9):2424–2432

[10] Najafi K, Wise K (1986) An implantable multielectrode array with on-chip signal processing. IEEE J Solid-State Circuits 21:1035–1044

[11] Nordhausen C, Maynard E, Normann R (1996) Single unit recording capabilities of a 100-microelectrode array. Brain Res 726:129–140

[12] Olsson RH III, Gulari M, Wise K (2002) Silicon neural recording arrays with on-chip electronics for in-vivo data acquisition. Paper presented at the 2nd annual international IEEE-EMBS special topic conference on microtechnologies in medicine and biology, Madison, Wisconsin, 2–4 May 2002

[13] Shenoy P, Miller K, Ojemann J, Rao R (2008) Generalized features for electrocorticographic BCIs. IEEE Trans Biomed Eng 55(1):273–280

[14] Steyaert M, Sansen W (1987) A micropower low-noise monolithic instrumentation amplifier for medical purposes. IEEE J Solid-State Circuits 22(6):1163–1168

[15] Wattanapanitch W, Fee M, Sarpeshkar R (2007) An energy-efficient micropower neural recording amplifier. IEEE Trans Biomed Circ Syst 1(2):136–147

Chapter 3
A Low-Power, Low-Noise, Open-Loop Amplifier for Neural Recording

The signal path in a neural recording system typically starts with an amplifier in order to boost the signal levels and buffer the high source impedance. Because of the small signal amplitudes, amplifier noise must be minimized in order to avoid unnecessary degradation of the signal. Additionally, the high impedance of neural electrodes necessitates a high impedance input.

For a fixed bandwidth, an amplifier's input-referred noise scales inversely with the square of its current consumption. In order to achieve acceptable noise levels, the front-end amplifier often consumes a substantial fraction of the overall system power [5]. Recently there has been a great deal of research into the design of low-power amplifiers for neural recording [3, 4, 9]. The large majority of previous work has focused on conventional closed-loop amplifiers built from operational amplifiers.

Op-amps in closed-loop configurations have been the work-horse of analog design for decades, due to the flexibility, precise gain, and linearity that can be achieved. Open-loop amplifiers have been used primarily in high-frequency applications, such as wireless design, where the loop gain needed to realize the benefits of a closed-loop architecture is difficult to attain. However, when power dissipation is a primary consideration, an open-loop topology may become attractive even for low-frequency applications.

Open-loop amplifiers can give superior noise performance for a given power budget at the expense of linearity performance, imprecise gain control, and reduced power-supply rejection. In this chapter, we will describe a simple open-loop amplifier design which achieved the lowest NEF to date.

3.1 Open-Loop Amplifier Design

The design philosophy behind the use of an open-loop amplifier is the idea that the unique nature of the of the neural recording problem justifies the acceptance of a penalty in linearity and supply rejection in exchange for maximum noise efficiency.

The small signal levels of neural signals relax linearity requirements relative to those for general purpose amplifiers. If the application is the detection of action potentials, then precise signal reconstruction is not as important as preservation of

J. Holleman et al., *Ultra Low-Power Integrated Circuit Design for Wireless Neural Interfaces*,
DOI 10.1007/978-1-4419-6727-5_3, © Springer Science+Business Media, LLC 2011

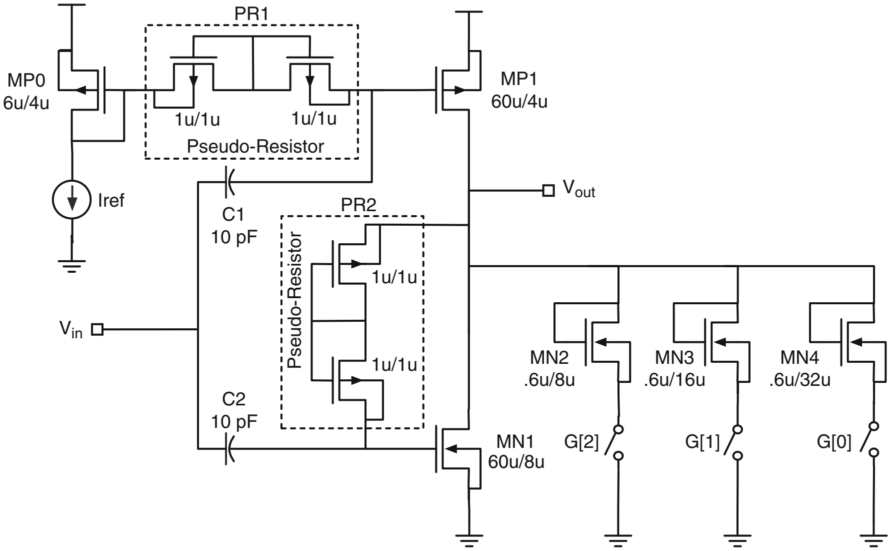


Fig. 3.1 Schematic of the open-loop amplifier

relative amplitudes, further relaxing both linearity requirements and the need for precisely defined gain. Provision of a stable power supply should be possible with careful system design. Implantation in the human body provides some shielding of the power supply against interferers such as 50/60 Hz noise. Low current consumption and low voltage requirements also ease the task of generating a stable supply.

A single-ended, open-loop amplifier designed for recording action potentials is shown in Fig. 3.1 [6]. MOS-bipolar pseudo-resistors (PR) [4] are used to implement the AC coupling necessary to reject large DC offsets due to contact potentials. Each of the transistors in the pseudo-resistor is connected such that there is a MOS diode and a parasitic source-bulk diode connected in anti-parallel. If the voltage across the device is small, then neither diode will conduct strongly, and the effective resistance is very large ($>10\,\mathrm{G}\Omega$). The voltage across PR1 is limited to the magnitude of the input signal, while the voltage across PR2 is a function of the output signal. In order to keep the pseudo-resistor in the high-resistance region, two devices are connected in series.

Two strategies are utilized here to minimize the input-referred noise for a given bias current. The first is to limit the number of current branches. There is only one branch operating at full current. The reference current is ten times smaller than the amplifier bias current, so it does not contribute significantly to the total power consumption. The same RC network used to AC couple the PMOS input presents a low-pass filter to the reference transistor MP0, so noise from the current reference is not added to the signal, permitting the use of a relatively noisy low-power bias generator.

The second strategy is to drive the gates of both MP1 and MN1. A conventional common-source amplifier has a current-source load which adds noise to the signal, but performs no amplification. Because the input must be AC-coupled, it is possible to decouple the DC levels of the gates of transistors MP1 and MN1 while keeping them connected in the frequency band of interest. The amplifier's transconductance is effectively doubled, while output noise remains constant, reducing the input-referred noise voltage spectral density by a factor of two. Because the bandwidth is determined by the load capacitor and is set based on the application requirements, the input referred RMS noise voltage is also reduced by a factor of two.

The aspect ratios of MP1 and MN1 were chosen to place both transistors in the weak inversion regime in order to maximize g_m/I_D. The lengths of the transistors MP1 and MN1 were chosen to be large to obtain sufficient gain from a single stage and to yield an acceptable level of $1/f$ noise, which is inversely proportional to gate area [1]. The bias current is generated from an on-chip bias circuit based on [2] and multiplied by a 3-bit digitally-controlled current mirror. The bias current in the amplifier can be varied from 110 to 770 nA.

This amplifier includes a bank of digitally-enabled diode-connected transistors M2–M4, which allow the user to control the gain through the gain-control word G[0:2]. The aspect ratio of MN2 is 100 times smaller than that of MN1, and $\mathrm{V}_{\mathrm{GS},1} = \mathrm{V}_{\mathrm{GS},2}$, so the incremental conductance of MN2 is approximately 100 times smaller than $g_{m,MN1}$. In the absence of any channel-length modulation, and assuming equal subthreshold slope factors (and thus equal g_m) for MN1 and MP1, MN2 would limit the gain to 200. This scheme was used to mitigate the risk of uncontrolled gain due to the open-loop topology. Including the effect of channel-length modulation, MN2 reduces the gain by about 6 dB, from 44.3 to 38.4 dB. With M3 and M4 enabled, the gain drops to 36.1 dB.

3.2 Results

This amplifier was fabricated in a 0.5 μm SOI-BiCMOS process, employing CMOS devices exclusively. It occupies 0.033 mm^2 and the current reference occupies an additional 0.013 mm^2 of die area. The entire circuit can operate from a supply between 1 V and 5 V, while the measurements presented here were taken with a 1.0 V supply.

Figure 3.2 shows the frequency response over the entire range of gain settings. The current reference is configured to provide the maximum bias current, yielding $\mathrm{I}_{\mathrm{DS}} = 770$ nA for MP1 and MN1. At the highest gain setting, the amplifier exhibits a gain of 44 dB and bandwidth of 1.9 kHz. The intermediate gain setting provides a gain of 38 dB and a 3-dB frequency of 3.6 kHz. With the lowest gain setting, the gain is 36 dB, and the 3-dB frequency is extended to 4.7 kHz. The remainder of this section will focus primarily on the low-gain setting, because it provides sufficient bandwidth to record action potentials. However, it is possible to extend the bandwidth at higher gain settings by increasing the bias current, either by overriding the internal bias generator, or with a modified design.

The input-referred noise spectrum of the amplifier is shown in Fig. 3.3. Despite the large transistor sizes, $1/f$ noise dominates. The total RMS noise at the input is 3.5 μV. It is difficult to discern the white thermal noise region of the spectrum because of the proximity of the $1/f$ noise corner to the output pole of the amplifier, but analysis predicts an input-referred thermal noise density of about 20 nV/$\sqrt{\mathrm{Hz}}$.

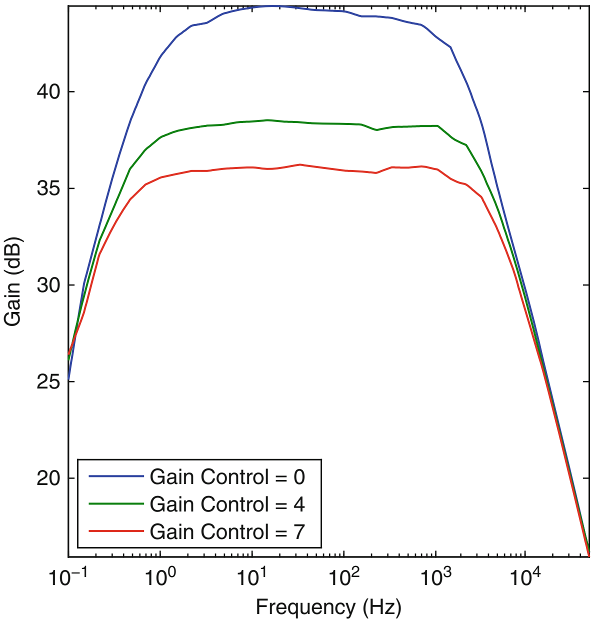


Fig. 3.2 The frequency response of the open-loop amplifier with three different gain settings. The gain adjustment number refers to the digital gain control word G[0:2] in Fig. 3.1

Feedback amplifiers achieve high linearity because their gain is determined by ratioed passive components. For open-loop amplifiers, nonlinearity of the transconductance and of the output impedance is manifested in a nonlinear input-output function. The linearity of the proposed amplifier can be assessed visually in Fig. 3.4a,

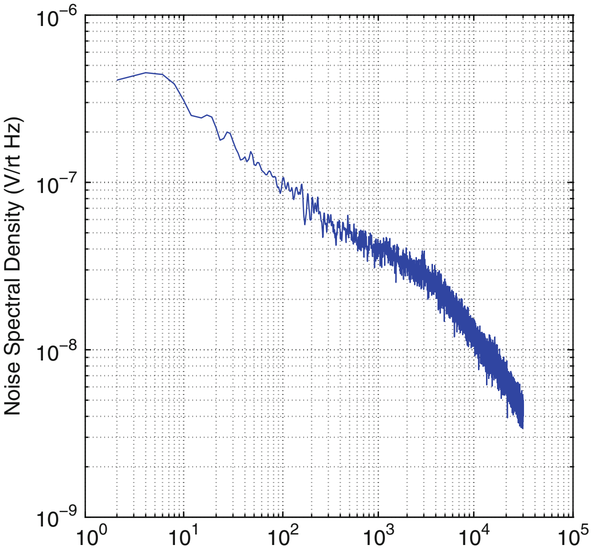


Fig. 3.3 Input-referred noise spectrum of the open-loop amplifier, computed as the measured output noise spectrum divided by the mid-band gain

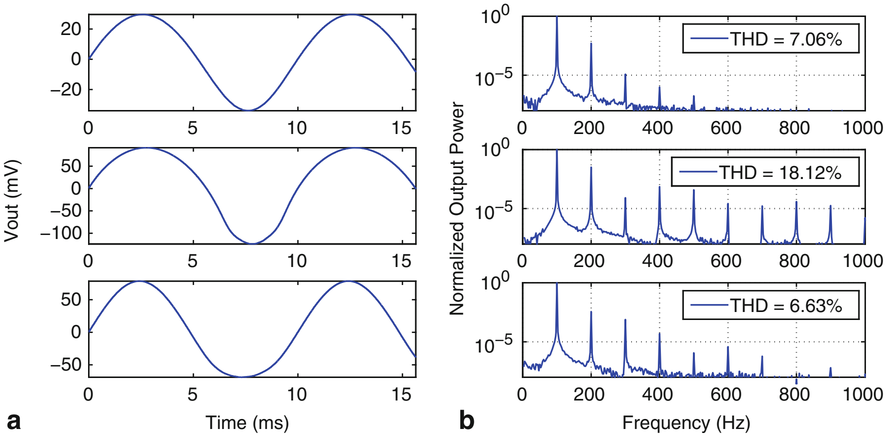


Fig. 3.4 **a** Output voltage waveforms for 100 Hz sinusoidal input. (Top) Low gain setting, peak-to-peak amplitude of 1 mV. (Middle) Low gain setting peak-to-peak amplitude of 10 mV. (Bottom) High gain setting, peak-to-peak amplitude of 1 mV. **b** The output power spectrum with the amplifier input driven by a 100 Hz sinusoid. The amplitude and gain are the same as in (**a**)

which shows output waveforms corresponding to a 100 Hz input with various amplitudes. In the top waveform, with peak-to-peak input amplitude of 1 mV, the distortion is not visually noticeable. With a 10 mV input, the incremental resistance of the gain-control transistors MN2-4 decreases at the upper end of the range, causing substantial compression. For the third waveform, the amplifier is in the high-gain configuration, and the input amplitude is 1 mVpp. Figure 3.4b shows the power spectra of the same three waveforms shown in Fig. 3.4a. THD with the 10 mV input is quite high at 18.12%, but for a 1 mVpp input, THD is lower, at 7.06 and 6.63% for the low and high gain settings, respectively.

In applications where a quiet power supply cannot be guaranteed, power-supply rejection ratio must be examined. In the proposed amplifier, both MP1 and MN1 have their sources connected to a power supply and their gates capacitively connected to the input. Thus, the positive and negative supplies directly modulate the P- and N-type transconductors, respectively. Therefore one would expect that the gain from the positive power supply to the output will be approximately half the gain from input to output, resulting in a minimal PSRR of 6 dB. Figure 3.5 shows the positive power-supply rejection ratio from 20 Hz to 20 kHz, which is an average of 5.5 dB between 1 and 100 Hz. Because of the weak supply rejection, the output will be susceptible to supply noise existing in the frequency band of interest.

Table 3.1 compares the performance of this amplifier to other published biosignal amplifiers. The noise efficiency factor (NEF), introduced in [7], is used to compare the noise and power performance to other amplifiers:

$$\mathrm{NEF} = V_{rms,in}\sqrt{\frac{2 \cdot I_{Total}}{\pi \cdot U_T \cdot 4kT \cdot BW}} \tag{3.1}$$

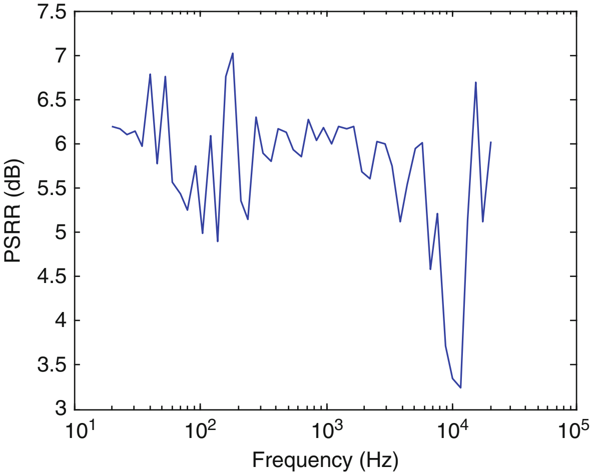


Fig. 3.5 Power-supply rejection ratio

Table 3.1 Comparison of neural amplifiers

	Gain (dB)	I_{Amp}	NEF	$v_{ni,RMS}$ (μV)	THD at Input (mVpp)	PSRR (dB)	Bandwidth
Harrison and Charles [4]	39.5	16 μA	4.0	2.2	1% at 16.7	≥85	0.025 Hz–7.2 kHz
Denison et al. [3]	45.5	1.2 μA	4.9	0.93	–	–	0.5–250 Hz
Wu and Xu [9]	40.2	330 nA	3.8	0.94	0.053% at 5	62	3 mHz–245 Hz
Wattanapanitch et al. [8]	40.2	330 nA	3.8	0.94	0.053% at 5	62	3 mHz–245 Hz
This work							
Open-loop	36.1	805 nA	1.8	3.6	7.1% at 1	5.5	0.3 Hz–4.7 kHz
Closed-loop	38.3	12.5 μA	2.48	1.95	1% at 1	63	0.023 Hz–11.5 kHz

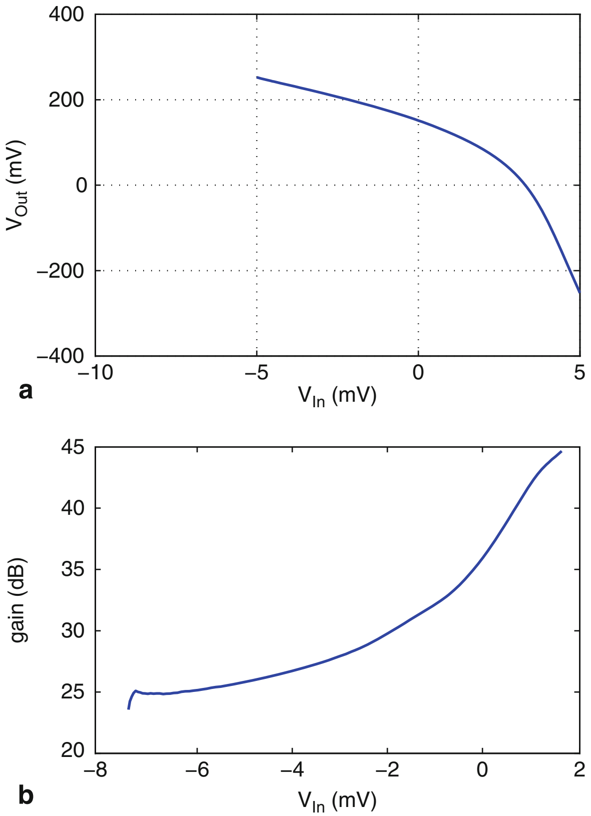
where I_{Total} is the total amplifier current, U_T is the thermal voltage, BW is the amplifier bandwidth, $V_{rms,in}$ is the input-referred RMS noise voltage.

For consistency with other work, the current specified in Table 3.1 excludes the current consumed by the bias generator, which consumes an additional 27 nA. This amplifier demonstrates the lowest NEF of any amplifier reported to date. Including the bias circuitry, the entire amplifier chip dissipates less than 1 μW.

3.3 Effect of Non-Linearity on Neural Recordings

As mentioned above, the open-loop amplifier achieves its noise efficiency at the expense of linearity. Figure 3.6a shows the input-output voltage relationship. The slope of this curve at any given point gives the amplifier's small signal gain for a

Fig. 3.6 **a** Voltage transfer curve for the open-loop amplifier. **b** Amplifier gain versus input voltage



signal centered at the corresponding input voltage, and is shown in Fig. 3.6b. Because the amplifier is AC-coupled, these curves cannot be measured at DC and are therefore constructed from sinusoidal input and output measurements.

To explore the effects of amplifier non-linearity on a variety of input signals, the data in Fig. 3.6a was fit to a polynomial, then various hypothetical input signals were amplified by the nonlinear gain represented by the polynomial fit.

Figure 3.7a shows a 150 µV spike amplified by the polynomial fit and imposed on the same spike amplified with perfect linearity. The two traces are nearly indistinguishable, suggesting that the effect of the amplifier's non-linearity is negligible for signals with the amplitude of a typical neural spike.

However, extracellular neural probes will also detect local field potentials (LFPs), which occur in the frequency band below 1 kHz and can have amplitudes as high as 5 mV [4]. Additionally, 50/60 Hz interference from wall power may corrupt the signals. If LFPs are in the amplifier's passband, they will shift any action potentials to a different point in the amplifier's input/output curve, effectively causing the relevant part of the input signal (the action potentials) to experience time-varying

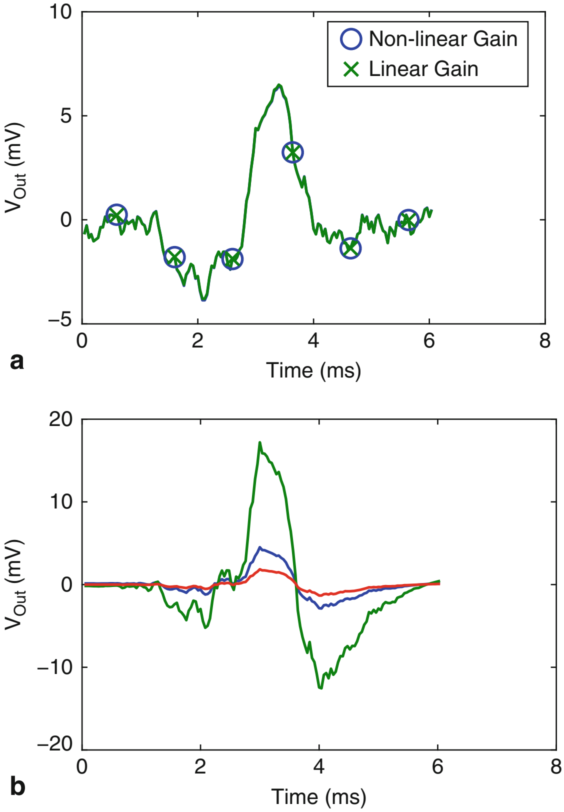


Fig. 3.7 **a** A 150 µV spike amplified with the nonlinear characteristic of the open-loop amplifier as well as with perfectly linear gain, showing that the amplifier nonlinearity is not significant for signals with the amplitude of typical neural spikes. **b** Three spikes amplified by the amplifier's nonlinearity after being superimposed on a much larger signal, representing interference due to LFPs or wall power

gain. In Fig. 3.7b, three spikes are added to a 5 mV sinusoid before being applied to the amplifier nonlinearity, to simulate the effect of a small spike being recorded in the presence of a large LFP interferer. In this case, the amplifier's nonlinearity does introduce significant errors, essentially subjecting the desired action potential signal to a time-varying gain.

A number of options are available to address this issue. A filter could be placed before the amplifier or built into the amplifier. For example, if the pseudo-resistors, with resistance of nearly 1 TΩ, were replaced with 16 MΩ resistances, the high-pass corner frequency would move to about 1 kHz. Local field potentials at 100 Hz would then be attenuated by 20 dB. If any large interferers are sufficiently attenuated by a filter, then we return to the situation shown in Fig. 3.7a, where the non-linearity does not have a noticeable effect. Another option is to compensate for the error in a later processing stage. The inverse non-linearity could be applied in the digital domain to recover the original signal, or spike detection and sorting algorithms which do not rely on amplitude information could be used. Finally, one could modify the amplifier to improve linearity. Chapter 4 describes an amplifier built with this strategy in mind.

3.4 Conclusions

This chapter describes two novel amplifiers which achieve excellent power efficiency. The open-loop amplifier exhibits the lowest NEF published to date, at the expense of linearity, supply rejection, and gain accuracy. The low power and area provided by this design would allow the realization of a 256-channel amplifier array with an area of 8.4 mm^2 and a power dissipation of 206 μW. While the noise efficiency of the closed-loop amplifier falls short of that achieved by the open-loop amplifier, it is still superior to that reported for any other closed-loop amplifier, demonstrating that the complementary-input topology used in both amplifiers is a powerful technique to improve noise-power efficiency.

The choice between a single-ended open-loop amplifier and a differential closed-loop amplifier depends on system-level considerations. The primary drawbacks of the open-loop LNA are gain inaccuracy, nonlinearity, and poor supply rejection. Because absolute amplitude is not typically a salient feature of neural recordings (due partially to other sources of amplitude uncertainty), the decision can be made based on linearity, supply rejection and power consumption considerations.

The supply rejection burden can be removed from the amplifier by using a regulator to provide a low-noise supply to all of the amplifiers in a system. In a multi-electrode system with many channels, the additional power consumption of the regulator is amortized across all of the channels. With only a small number of channels, the additional power per channel for the regulator may be greater than the power saved by using a single-ended topology. Thus, the single-ended topology becomes more attractive for higher channel-count recording systems.

It should also be noted that the three strategies employed in this amplifier—complementary input drive, open-loop topology, and single-ended input stage—are essentially unrelated. Any amplifier can be operated in an open-loop configuration. Conversely, the single-ended amplifier could be combined with an additional gain stage and a capacitive feedback network to improve linearity. The complementary input drive can also be applied to a closed-loop or differential design, as will be demonstrated in Chap. 4.

References

[1] Allen P, Holberg D (2002) CMOS analog circuit design. Oxford University Press, New York

[2] Camacho-Galeano E, Galup-Montoro C, Schneider M (2005) A 2-nW 1.1-V self-biased current reference in CMOS technology. IEEE Trans Circ Syst II: Express Briefs [see also IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing] 52(2): 61–65

[3] Denison T, Consoer K, Kelly A, Hachenburg A, Santa W (2007) A 2.2 μW 94 nV/pHz chopper-stabilized instrumentation amplifier for EEG detection in chronic implants. In: international solid-state circuits conference, Digest of Technical Papers, San Francisco

[4] Harrison R, Charles C (2003) A low-power low-noise CMOS amplifier for neural recording applications. IEEE J Solid-State Circuits 38(6):958–965

[5] Harrison R, Watkins P, Kier R, Lovejoy R, Black D, Greger B, Solzbacher F (2007) A low-power integrated circuit for a wireless 100-electrode neural recording system. IEEE J Solid-State Circuits 42(1):123–133

[6] Holleman J, Otis B (2007) A sub-microwatt low-noise amplifier for neural recording. In: 29th annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBS 2007), Lyon, pp 3930–3933

[7] Steyaert M, Sansen W (1987) A micropower low-noise monolithic instrumentation amplifier for medical purposes. J Solid-State Circuits 22(6):1163–1168

[8] Wattanapanitch W, Fee M, Sarpeshkar R (2007) An energy-efficient micropower neural recording amplifier. IEEE Trans Biomed Circuits Syst 1:136–147

[9] Wu H, Xu Y (2006) A 1 V 2.3 μW biomedical signal acquisition IC. In: Proceedings of the 2006 IEEE international conference on solid-state circuits, Digest of Technical Papers, San Francisco, CA, pp 119–128

Chapter 4
Closed-Loop Neural Recording Amplifier Design Techniques

As previously described, in order to accommodate weak neural signals, we need sufficient amplification and signal conditioning at the front-end of a neural-recording system. Specifically, the requirements on the front-end amplifier can be summarized as below:

- Input-referred noise voltage $<10\,\mu$V
- Midband gain $\cong$40 dB
- Input impedance $\geq$ a few MΩs at 1 kHz
- Pass-band compatible with the desired signals (see Table 2.1).
- AC-coupled input in order to block DC offsets.
- Small silicon area and no off-chip components.
- CMRR, PSRR $\geq$60 dB
- Power dissipation $\ll$100 μW/channel

In this chapter, the design methodologies of two new closed-loop amplifier architectures are presented. The implementation details of these topologies are then compared and contrasted. Afterward, we will discuss the design of a variable-gain amplifier with six variable gain settings from 0 to 40 dB, making the amplifiers suitable to process a variety of signals.

4.1 Design of a Closed-Loop Telescopic Amplifier

4.1.1 Closed-Loop Architecture

To increase the output signal swing with low supply voltages (as low as 1V), we chose a fully-differential topology. We also chose a closed-loop topology in order to improve the CMRR, PSRR, linearity, and gain precision. Figure 4.1 shows the closed-loop architecture.

The input signals are AC-coupled into the amplifier to reject large DC offsets from the electrode-tissue interface, preventing saturation at the output of the amplifier. The sizing of the input capacitors, C_s, also need to be carefully chosen. They must be

J. Holleman et al., *Ultra Low-Power Integrated Circuit Design for Wireless Neural Interfaces*,
DOI 10.1007/978-1-4419-6727-5_4, © Springer Science+Business Media, LLC 2011

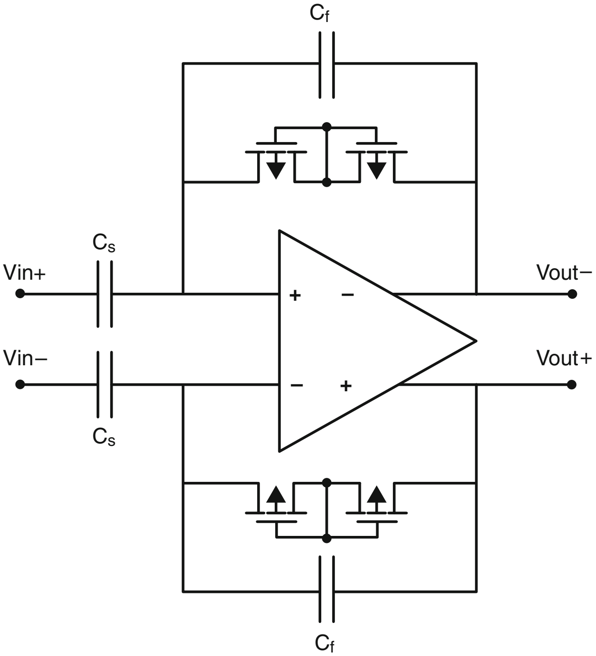


Fig. 4.1 Closed-loop amplifier schematic

made small enough (the input impedance, $\frac{1}{j\omega C}$ is large enough) to avoid attenuation of the input signal from the electrode, but large enough to avoid attenuation from the capacitive divider it forms with the parasitic capacitance of the input devices. As shown in (4.1), the capacitive divider also increases the input-referred noise of the LNA. Let $\overline{v_{ni,amp}{}^2}$ and $\overline{v_{ni}{}^2}$ represent the input-referred noise of the OTA and the LNA, respectively. The mid-band gain is set by C_s/C_f.

$$\overline{v_{ni,amp}{}^2} = \left(\frac{C_s + C_f + C_{in}}{C_s}\right)^2 \overline{v_{ni}{}^2} \tag{4.1}$$

4.1.2 Analysis of Pseudo-Resistors

Because signals are AC-coupled into the amplifier, the gates of the input transistors need to be properly biased to ensure proper operation. At the same time, in order to pass EEG and LFP signals in the sub-Hertz band, we need to form a sub-Hertz high-pass frequency corner. To address both design concerns, we would need to use large resistors that would normally take up significant chip area. Here, we chose to use pseudo-resistors [2] to bias the gates to avoid large resistors. Pseudo-resistors (Fig. 4.2) are MOS-bipolar devices which have equivalent resistance in the order of

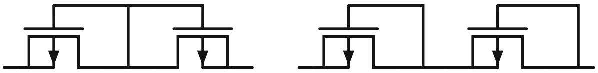


Fig. 4.2 Two configurations of pseudo-resistors

100 GΩ to 1 TΩ if the voltage drop across them is small enough (<0.2 V, according to [2]). When $V_{GS} > 0$, the parasitic source-well-drain p-n-p junction acts as a diode-connected BJT; when When $V_{GS} < 0$, each device functions as a diode-connected pMOS transistor.

We cascaded two long-channel (50 µm), minimum-width pMOS transistors to increase the equivalent resistance of the pseudo-resistors and to ensure sufficiently large resistance with reasonable voltage drop across them. The large incremental resistance also takes into account the drop in resistance in the presence of large input amplitudes. The low-frequency cutoff ω_L of the amplifier is then $\frac{1}{2r_{inc}Cf}$, with r_{inc} being the incremental resistance of the pseudo-resistor. However, this frequency corner is difficult to know a-priori as the incremental resistance at the parasitic source-well-drain p-n-p junction is poorly modeled. This sub-Hertz corner creates a large time constant, resulting in slow start-up and settling time of the amplifier. There are other possible replacement for pseudo-resistors. For instance, some bioamplifier designs have used transistors biased in the subthreshold region to approximate large-valued resistors [1]. This technique functions similarly while requiring additional biasing circuitry.

4.1.3 Telescopic OTA Design Overview

In order to lower the power consumption and ease integration with complex digital subsystems, the amplifier, shown in Fig. 4.3, should operate from a supply as low as 1 V. Although neural-recording applications can tolerate gain error, linearity requirements dictate high open-loop gain. A single-stage amplifier topology does not provide sufficient gain or output swing with a 1 V supply. Therefore, we chose to employ a two-stage fully-differential design to simultaneously satisfy the gain

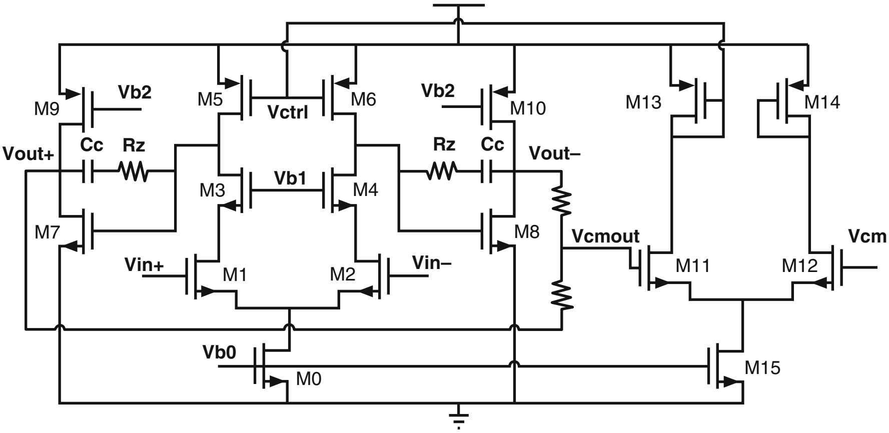


Fig. 4.3 Telescopic OTA schematic

and swing requirements. Common-mode feedback is included to stabilize the output common-mode voltage.

The design of the first stage is crucial in achieving an optimal power-noise tradeoff while meeting gain specifications. Although the design in [2] optimized the power-noise tradeoff nearly to the theoretical limit of its particular OTA topology, the topology itself is not optimized for power because a large portion (50%) of current is consumed in the biasing branch. Although [5] optimized the power-noise tradeoff for a folded-cascode amplifier, there is no clear advantage to using a folded-cascode topology in the first-stage. Although the folded-cascode topology increases the input common-mode range, this is typically not a concern since the low-level input signal is AC-coupled into the amplifier anyway. Furthermore, the folded-cascode topology generally consumes more current and has worse noise performance because of the extra current branches. Therefore, a telescopic-cascode topology is used in the first stage to achieve sufficient gain while conserving current. To accommodate the low supply voltage, we have omitted the pFET cascode transistors, but retained the nFET cascodes to provide input-output isolation and sufficient first-stage gain.

The second-stage uses a standard common-source topology to implement a gain stage. The tail current source in the second stage is eliminated to ensure sufficient voltage swing under low supply voltages.

4.1.4 Design Optimization

Transistor sizing is critical for simultaneously achieving low noise and low power. We chose 6 µA for the bias current I_{bias}, giving $M_1 - M_8$ 3 µA.

Table 4.1 shows the parameters and operating conditions of each transistor in the OTA. Neglecting the noise contribution of the second stage, the input-referred thermal noise power can be simplified to

$$V_{ni,th}^2 = \left(\frac{16kT}{3g_{m1,2}}\left(1+\frac{g_{m5,6}}{g_{m1,2}}\right)\right)\Delta f \tag{4.2}$$

From this equation, we want to have $g_{m5,6} \ll g_{m1,2}$ to minimize the thermal noise contribution. This can be accomplished by sizing $(W/L)_{5,6} \ll (W/L)_{1,2}$. Thus, the NMOS input pair $M_{1,2}$ is pushed to weak inversion, where the g_m/I_D ratio is maximized. By sizing the input pair $M_{1,2}$ to 624/2 µm/µm, the input transistors operate in the deep subthreshold region, and the g_m/I_D ratio is maximized. The large

Table 4.1 Device parameters of telescopic-cascode closed-loop biopotential amplifier

	W/L (µm)	I_d (µA)	Inv. Coeff	g_m/I_d (V^{-1})	$\lvert V_{GS} - Vt\rvert$ (mV)
$M_{1,2}$	616/2	3	0.023	27.56	154
$M_{3,4}$	12/5	3	0.27	22	101
$M_{5,6}$	12.2/13	3	11.3	6.9	257.7
M_0	109.8/8	6	0.54	19.4	386
$M_{7,8}$	12/4	2.2	0.96	16.82	32
$M_{9,10}$	8.6/3	2.2	1.92	13.64	106

gate area also reduces 1/f noise, as discussed in the next paragraph. At the same time, by sizing PMOS load $M_{5,6}$ to 12.2/26 µm/µm, the transistors are pushed to strong inversion, where the g_m/I_D ratio is minimized. The sizing of cascode transistors M_2 is not crucial because the noise of the cascode transistors negligibly contributes to the output, especially at low frequencies [4]. Therefore, we have sized it to be in moderate inversion. However, the total capacitance C_x at the source of the cascode transistors gives rise to the contribution of the noise of the cascode transistors at high frequency (specifically, higher than $g_{m2}/2\pi C_x$). In addition to the increase in the noise, the gain is also decreased as C_x shunts the signal current produced by input pair $M_{1,2}$ to ground. However, since the frequency $g_{m2}/2\pi C_x$ is farther out than the bandwidth of the amplifier, we could ignore these secondary high frequency effect on the noise and gain of the first-stage OTA.

The input devices should be large to reduce their flicker noise. However, large input devices increase the input parasitic capacitance C_{in}, which attenuates the signal at the OTA input via the capacitive divider (4.1). An optimization balancing these tradeoffs led to a sizing of 616/2 µm/µm for the input devices.

4.1.5 Stability and Common-Mode Feedback

Two-stage amplifier design necessitates a compensation capacitor to split the poles at the output of the first and second stage. In order to eliminate the feed-forward zero, we added a nulling resistor with a value of approximately $1/g_m$ of the input transistor of the second stage. For this high gain fully-differential amplifier, an internal common-mode feedback (CMFB) path must be added to establish a common mode output voltage over the frequencies of interest. We chose to sense the common-mode output voltage using two large resistors and generate a continuous-time CMFB control signal back into the differential mode path.

The signal propagation of the common-mode feedback path consists of two parts: from the average output V_{cmout} to the feedback control V_{ctrl}, and from V_{ctrl} to the amplifier output. The first part of the CMFB path has a wide bandwidth and small DC gain ($\simeq 1$); the second part determines the CMFB frequency response as illustrated in Fig. 4.4. Let $g_{m1-4,6-8}$ denote the transconductance of the corresponding transistors, C_c and C_L denote the compensation and load capacitors, g_{o1} and g_{o2} denote the total output conductances of stage 1 and 2. Then

$$A_{cmfb} = \frac{V_{out,CM}}{V_{ctrl}} = \frac{-sg_{m5,6}C_c + g_{m5,6}g_{m7,8}}{s^2C_cC_L + sC_cg_{m7,8} + g_{o1}g_{o2}} \tag{4.3}$$

$$A_{dm} = \frac{V_{out,DM}}{V_{in,dm}} = \frac{-s(g_{m1,2})C_c + (g_{m1,2})g_{m7,8}}{s^2C_cC_L + sC_cg_{m7,8} + g_{o1}g_{o2}} \tag{4.4}$$

Both the differential and common-mode gains share the same compensation capacitor C_c and $g_{m7,8}$ stage. The similarity of the topologies leads to a stable CMFB path if the differential-mode path is unity-gain stable. The CMFB gain and bandwidth must be larger than those of the common-mode path. Although the requirement on

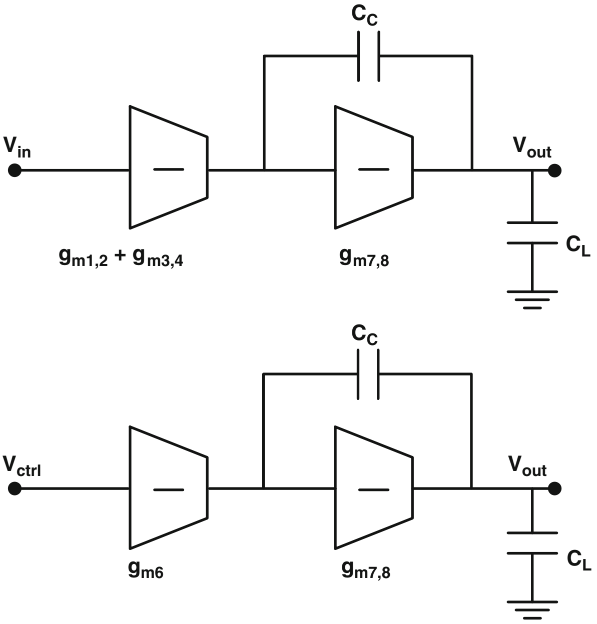


Fig. 4.4 Top: Differential-mode gain path. Bottom: CMFB gain path

CMFB gain and bandwidth is relaxed as the common-mode path has small gain and bandwidth, a high CMFB gain is preferred to achieve more accurate common-mode voltage, and a high CMFB bandwidth is preferred to improve the CMRR at high frequencies. This CMFB topology achieves both high gain and bandwidth, while saving power by sharing one CMFB circuit between both the first and second stages.

4.2 Design of a Closed-Loop Complementary-Input Amplifier

In the last section, we discussed in detail the design of a low-noise low-power amplifier using conventional telescopic-cascode technique. In this section, we will introduce a technique that eases the trade-off between power consumption and noise performance by employing a complementary-input strategy. In the discussion below, "LNA1" will refer to the closed-loop telescopic-cascode amplifier, and "LNA2" refers to the closed-loop complementary-input amplifier.

4.2.1 Design of an Closed-Loop Fully-Differential Complementary-Input Amplifier

AC coupling at the inputs using 20 pF capacitors and high-resistance MOS-bipolar pseudoresistors prevent offset amplification, similar as the previous section. Thick-oxide MOS transistors with large gate areas are used at the input to reduce gate leakage while minimizing 1/f noise.

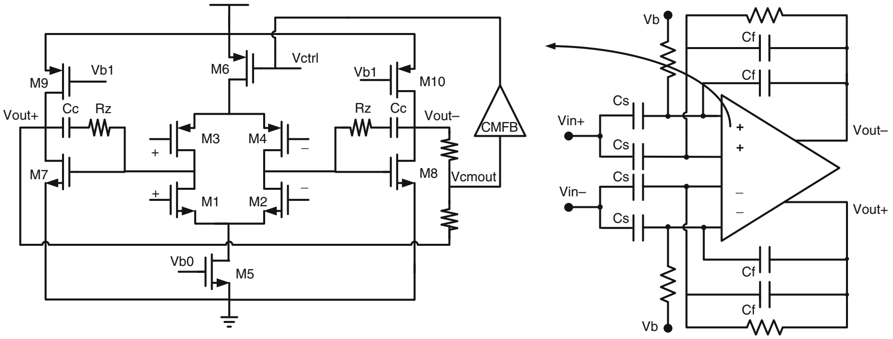


Fig. 4.5 Complementary amplifier schematic [3]

As shown in Fig. 4.5, input capacitors C_s separate the signal from the bias path, which allows the input to simultaneously drive the n- and pFET transistors of the first stage. Similar to its open-loop counterpart, the input-referred noise voltage is reduced by a factor of $\sqrt{2}$. The input-referred thermal noise power is twice that of the single-ended design because differential branches double the output noise. If $g_{m1} = g_{m3}$, it can be expressed as:

$$V_{ni,th}^2 = \left(\frac{16kT}{3g_{m1} \cdot 2}\right) \Delta f \tag{4.5}$$

Similarly, (4.2) shows the input-referred noise for LNA1. If $g_{m5} = 0$, than (4.2) reduces to

$$V_{ni,th}^2 = \left(\frac{16kT}{3g_{m1}}\right) \Delta f \tag{4.6}$$

A comparison between (4.5) and (4.6) reveals that the input-referred noise voltage of LNA2 is approximately $1/\sqrt{2}$ that of LNA1.

Determining the appropriate level of inversion is crucial in minimizing noise and power. Similar to LNA1, 6 µA is chosen for the bias current in M_0, giving $M_1 - M_8$ drain currents of 3 µA. Table 4.2 shows the parameters and operating conditions of each transistor in the OTA. Common-centroid techniques are used for the input transistors M_{1-4}. By sizing both input pairs to 552/2 µm/µm, the input transistors

Table 4.2 Device parameters of complementary-input closed-loop biopotential amplifiers

	W/L (µm)	I$_d$ (µA)	Inv. Coeff	g_m/I_d (V^{-1})	$\lvert V_{GS} - Vt\rvert$ (mV)
$M_{1,2}$	552/2	3	0.022	27.53	152
$M_{3,4}$	552/2	3	0.12	24.23	106
M_5	110.4/8	6	0.54	19.4	5
M_6	73.2/8	6	1.98	13.5	103
$M_{7,8}$	8.6/3	2.1	1.93	13.62	103
$M_{9,10}$	12/4	2.1	0.93	16.96	30

are operating in deep subthreshold region where g_m/I_D is maximized. The nFET and pFET input transistors have the same aspect ratio so that both signal paths experience the same attenuation through the input capacitive divider formed by C_s, C_f, and parasitic capacitance (4.1). The current source transistors have long lengths (8 µm) to allow a high output impedance. The noise from the current source appears as common-mode noise and is largely rejected by the differential operation.

Because we are driving both NMOS and PMOS inputs, we also need two capacitive feedback paths from the output to the gates of both NMOS and PMOS input transistors. The closed-loop gain is also set to C_s/C_f. Similar to the previous amplifier, C_s is set to 20 pF, and C_f is set to 200 fF for a closed-loop gain of 40 dB. In addition to the differential input stage, we also added common-source second stage to increase the open-loop gain, decrease the noise contribution from the second stage, as well as maximizing the output swing under low-supply-voltage conditions.

Because the input also drives the PMOS transistor pair $M_{3,4}$, the transconductances of $M_{3,4}$ not only contribute to the differential gain, but also the common-mode gain. Without PMOS tail source M6, the common-mode gain would be approximately half of the differential gain, mainly contributed by the signal path through the PMOS input. In order to ensure high CMRR, we use dual tail current sources in the first stage to degenerate the common-mode transconductance, thus reducing the common-mode gain. This configuration also improves the PSRR. Any variation in the supply is attenuated by approximately $\frac{g_{m6}}{(g_{m3}+g_{m4})} \cdot (1 - \frac{V_{g6}}{V_{dd}})$ before being amplified by the g_m mismatches in $M_{3,4}$ (4.8). This power supply gain is lower than that of a conventional telescopic-cascode amplifier, where any variation in the supply is directly amplified by the g_m mismatches in the PMOS load transistors. Let $g_{o1,2}$ denote the output conductance of the first and second stage, $g_{o5,6}$ denote the output conductance of current source transistors $M_{5,6}$, g_{m8} denote the transconductance of the second stage, Δg_m denote the g_m mismatch in $M_{3,4}$, and C_c denote the compensation capacitor. The common-mode gain (A_{cm}) and the gain of power-supply interference (A_{ps}) can be expressed as

$$A_{cm} = \frac{V_{out}}{V_{in,cm}} \simeq \frac{(g_{o5} + g_{o6})g_{m8}/(g_{o1}g_{o2})}{1 + sC_c/(g_{o5} + g_{o6})} \tag{4.7}$$

$$A_{ps} = \frac{V_{out}}{V_{in,supply}} \simeq \frac{\Delta g_m \gamma g_{m8}/(g_{o1}g_{o2})}{1 + sC_c/\Delta g_m} \Bigg| \gamma = \frac{g_{m6}}{(g_{m3} + g_{m4})} \cdot \left(1 - \frac{V_{g6}}{V_{dd}}\right) \tag{4.8}$$

Similar to LNA1, the fully-differential topology necessitates common-mode feedback to stabilize the output common-mode voltage. The output common-mode voltage sensed by resistors is compared with a reference voltage in a single-stage differential amplifier. The output of the CMFB amplifier is fed back to control the gate voltage of M_6, thus adjust the output of the first stage. Similarly, this feedback topology allows compensation capacitors and resistors to be shared between the CMFB path and differential-mode path in order to ensure a high CMFB gain and a large CMFB bandwidth. The common-mode gain expression is the same as in LNA1.

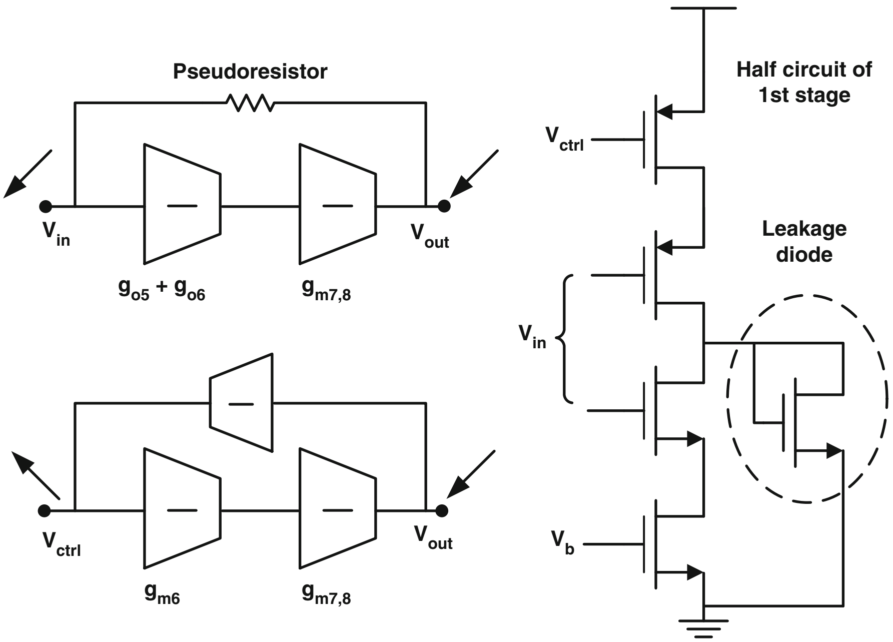


Fig. 4.6 Closed-loop amplifier start-up concern alleviated by adding a leakage path through a diode-connected transistor at the first-stage output

In order to provide DC feedback and bias the input transistors, the outputs are fed back through pseudoresistors to bias the gates of the NMOS input transistors $M_{1,2}$. However, this feedback inevitably forms a positive feedback loop at low frequencies. As shown in Fig. 4.6, this is particularly problematic when output common-mode voltage is initially low. In this case, the pull-down paths are turned off as the gates of $M_{1,2}$ are low. At the same time, the common-mode feedback control voltage rises, which also turns off the pull-up paths, leaving the first stage output in a high-impedance state. To ensure reliable start-up, we added a pair of diode-connected transistors at the output of the first stage connecting to ground. This scheme provides additional current paths through the diode-connected transistors when both the pull-up and pull-down paths are initially turned off. The additional currents are small enough that they do not affect the normal operation of the amplifier.

4.3 Design of a Variable-Gain Amplifier

In order to adapt to input signals of different amplitudes (500 µV–5 mV), a variable-gain amplifier (VGA) is added after the front-end amplifier. Because this is the second stage, its noise is attenuated by the first-stage amplifier gain (40 dB). Therefore, the input-referred noise requirement on the VGA is not stringent. As such, we will not

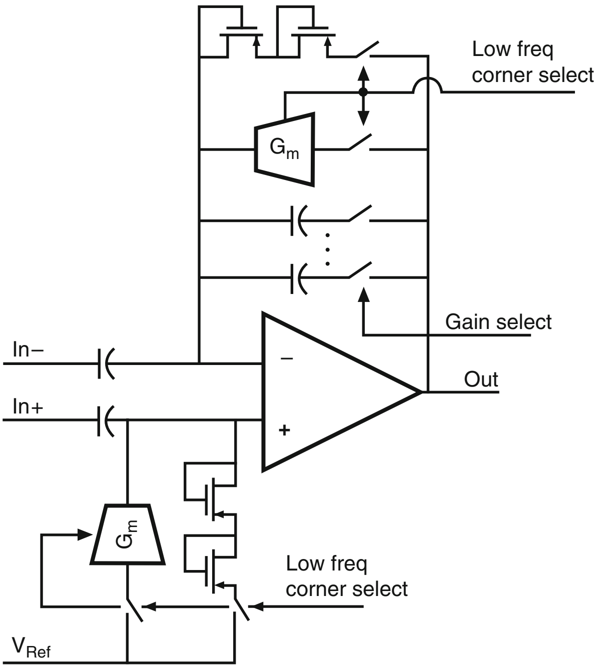


Fig. 4.7 VGA closed-loop schematic [3]

focus on the power-noise optimization for the VGA but will instead discuss the necessary programmability of the VGA.

The VGA consists of a complementary rail-to-rail folded-cascode core to improve the input signal swing. As shown in Fig. 4.7, a capacitor array consisting of six sets of capacitors and switches is placed in the feedback path. The six variable gains of the VGA are programmable by selecting any one of the six feedback capacitors. The capacitors are selected so that the closed-loop gain logarithmically spans from 0–38 dB. In addition to the selectable gain settings, we have incorporated the adjustable low-frequency high-pass corner so that the corner can stay relatively constant across the various gain settings. This high-pass corner not only prevents propagation of DC offsets at the output of the LNA, but also rejects the low-frequency interference (i.e., 60 Hz). The six variable low-frequency corners are set by programming the feedback transconductor bias current. The feedback transconductors are standard five-transistor g_m cells, which current can be changed by mirroring over different fractions of the bias current. The current in the g_m cells can be as small as several nanoamps. Alternately, pseudoresistor feedback can be selected to obtain a low-frequency corner below 10 Hz. This is helpful in amplifying LFP, ECoG, or EMG signals that have useful signals below 10 Hz. An on-chip shift register can shift in the proper configuration bits to enable the different gain and frequency settings.

References

[1] Chandran A, Najafi K, Wise K (1999) A new DC baseline stabilization scheme for neural recording microprobes. In: Proceedings of the first joint BMES/EMBS conference, annual fall meeting of the Biomedical Engineering Society, 21st annual international conference of the Engineering in Medicine and Biology Society, Atlanta, Georgia, pp 386–387

[2] Harison R, Charles C (2003) A low-power, low-noise CMOS amplifier for neural recording applications. IEEE J Solid-State Circuits 38:958–965

[3] Rai S, Holleman J, Pandey J, Zhang F (2009) A 500 µW neural tag with 2 µVrms analog front-end and frequency multiplying MICS/ISM FSK transmitter. In: IEEE international solid-state circuits conference, Digest of Technical Papers, pp 212–213

[4] Razavi B (2000) Design of analog CMOS integrated circuits. Tata McGraw-Hill Edition, India

[5] Wattanapanitch W, Fee M, Sarpeshkar R (2007) An energy-efficient micropower neural recording amplifier. IEEE Trans Biomed Circuit 1(2):136–147

Chapter 5
Closed-Loop Bio-Signal Amplifiers: Experimental Results

In this chapter, we will present the measurement results of the telescopic-cascode and complementary low-noise amplifiers discussed in the previous chapter. In order to compare and contrast the performance of these two LNA designs, we fabricated both LNAs in a 0.13 μm CMOS process. As the analog front-end of a neural-recording channel, each LNA is followed with a variable-gain amplifier (VGA) to accommodate signals of various amplitudes. Figure 5.1 illustrates one recording channel with the complementary LNA followed by a VGA.

Figure 5.2 shows the layout of the analog front-end (AFE) section of the system. Metal-Insulator-Metal (MIM) or dual MIM capacitors are mainly used for their small area/capacitance, good linearity and low substrate capacitance. The conventional amplifier design (top) uses 46,800 μm^2 of silicon and, 57.8% of this area is taken up by capacitors. The complementary amplifier design (bottom) uses 71,750 μm^2 of silicon and, 67.4% of this area is taken up by capacitors.

5.1 Amplifier Testing

In this section, we will compare and contrast the two different LNA designs: the telescopic and complementary fully-differential amplifiers (referred to as LNA1 and LNA2, respectively).

Figure 5.3a compares the frequency response of LNA1 and LNA2. The mid-band gain of LNA1 is 40.5 dB, whereas that of LNA2 is 40 dB. The minute difference is likely due to the different layout of the feedback capacitors. In the LNA2, we tied the bottom plates of two feedback capacitors that connect to the same output, resulting in a higher effective C_{fb}. The closed-loop gain is thus slightly lower than that of the LNA1. The −3 dB low-pass corners occur at approximately 8 kHz for LNA1, and 10 kHz for LNA2. The difference can be attributed to the larger effective transconductance G_m of LNA2. The low-frequency high-pass corner of LNA2 (0.05 Hz) is lower than that of LNA1 (0.4 Hz) because a longer length of pseudoresistor is used for LNA2 design.

Figure 5.3b compares the input-referred noise spectrum of LNA1 and LNA2. Consistent with theory, the input-referred noise of LNA1 is higher than that of LNA2

J. Holleman et al., *Ultra Low-Power Integrated Circuit Design for Wireless Neural Interfaces*,
DOI 10.1007/978-1-4419-6727-5_5, © Springer Science+Business Media, LLC 2011

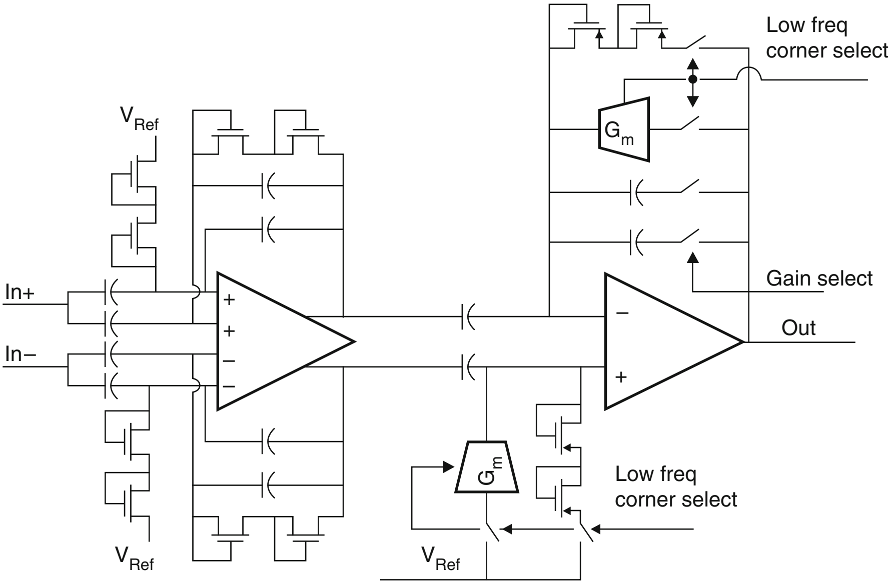
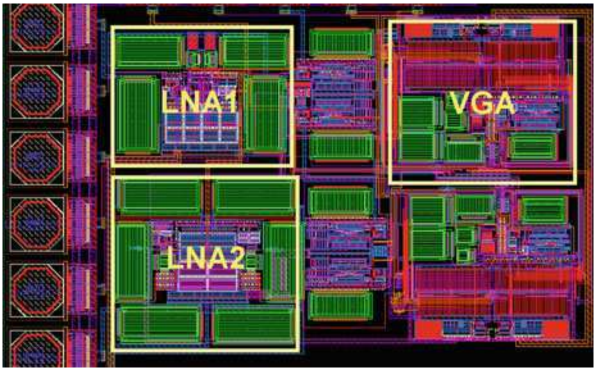


Fig. 5.1 One neural-recording channel with complementary amplifier design

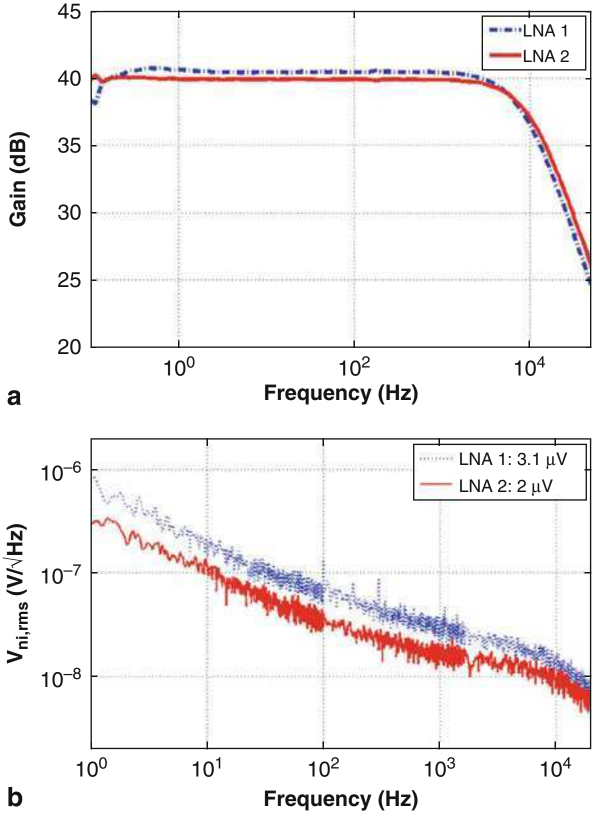
Fig. 5.2 Layout of the bio-signal interface circuit



because of the larger effective G_m in LNA2. We can also observe that the flicker noise dominates the entire frequency range of interest. The measured input-referred noise integrated from 0.1 Hz to 25 kHz are 3.1 µV, 3.5 µV, and 2 µV, respectively.

Figure 5.4a compares the PSRR of the two LNAs. The PSRR for LNA1 is approximately 20 dB lower than that of LNA2. This is consistent with the discussion earlier, where the PSRR of LNA2 is improved by the ratio $\frac{g_{m6}}{(g_{m3}+g_{m4})} \cdot (1 - \frac{V_{g6}}{V_{dd}})$. Superior PSRR performance can also be attributed to better transistor device matching in LNA2 compared with LNA1. Figure 5.4b compares the CMRR of LNA1 and LNA2. The CMRR for LNA1 has an average value of 60 dB, compared with 80 dB for LNA2. The larger devices in LNA2 should result in smaller expected values of CMRR due to reduced mismatch.

Fig. 5.3 **a** Bode magnitude and phase plots comparing LNA1 with LNA2; **b** Noise plot comparing LNA1 with LNA2



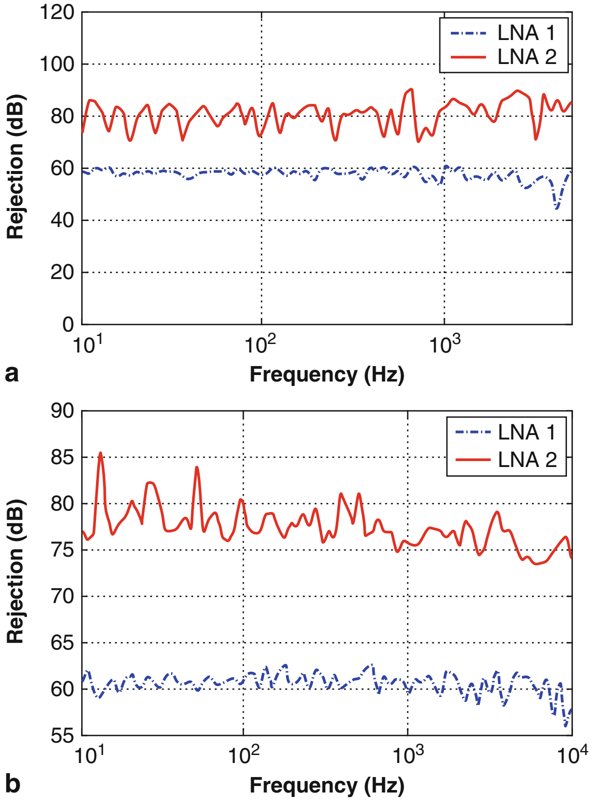
Finally, the linearity of the amplifiers is examined. Many papers use total harmonic distortion (THD) to describe linearity. However, in our experience, the main concern for spike-recording applications is gain compression due to interferers such as electromagnetic interference or low frequency local field potentials that can result in time-varying gain. Therefore, it is more useful to characterize the −1 dB gain compression point (approximately 89% of voltage gain, or 80% of power gain) than THD. We will evaluate the linearity performance by comparing their −1 dB gain compression input voltage. As shown in Fig. 5.5, the −1 dB gain compression point occurs at input level of 3 mV for LNA1, and 4 mV for LNA2. The difference can be attributed to the complementary-input topology employed in LNA2.

The two amplifiers are compared with other neural amplifiers in Table 5.1.

5.2 Variable Gain Amplifier (VGA) Testing

The measured gain settings are varied from 3 to 38 dB by adjusting the gain-control bits. Figure 5.6a shows the total gain of the LNA–VGA chain. In this plot, the high-pass frequency corners are adjusted according to the gain setting so that the corner

Fig. 5.4 **a** PSRR bode magnitude comparison of the two LNAs; **b** CMRR magnitude comparison of LNA1 and LNA2



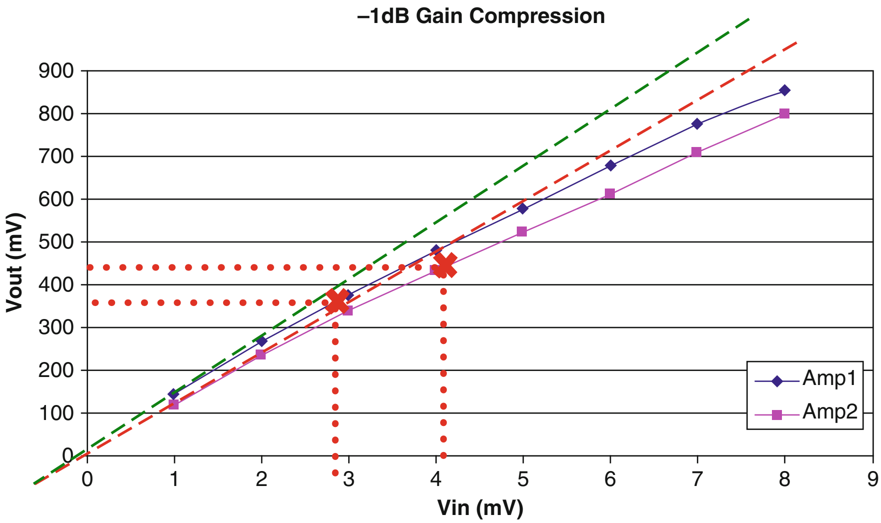


Fig. 5.5 −1 dB gain compression of LNA1 compared with LNA2

Table 5.1 Performance comparison of biopotential amplifiers

	LNA1	[3]	LNA2	[2]	[1]	[4]	[5]
Vdd (V)	1	1	1	+/−2.5	1.8–3.3	2.8	0.8–1.5
I_{Amp} (μA)	12.5	0.8	12.1	16	1.2	2.7	0.33
NEF	4.4	1.8	2.6	4.0	4.9	2.67	3.8
Gain (dB)	40.5	36	40	39.5	45.5	30.8	40.2
1 dB comp. at Input (mV)	3	1.7	4	–	–	–	–
$v_{ni,RMS}$ (μV)	3.1	3.6	2	2.2	0.93	3.06	2.7
PSRR (dB)	≥60	5.5	≥80	≥85	–	75	62–63
Bandwidth (Hz)	0.4–8.5 k	0.3–4.7 k	0.05–10.5 k	0.025–7.2 k	0.5–180	45–5.3 k	3 m–245
Area (mm^2)	0.047	0.046	0.072	0.16	–	0.16	1
Technology (μm)	0.13	0.5	0.13	1.5	0.8	0.5	0.35

frequencies stay relatively constant (∼300 Hz) across all gain settings. The dashed lines are the corresponding simulated bode plots. The realized frequency corners are higher because more current is sourced in the g_m cells that set the VGA's high-pass corner. The measured high-pass corner varied from below 10 Hz (pseudoresistor setting) to 400 Hz by shifting in different control values. Figure 5.6b is plotted at the lowest VGA gain setting. Based on the input signal characteristics, we can choose different VGA gain and high-pass corner frequency settings.

Figure 5.7 shows the measured input-referred voltage noise spectrum of the LNA itself and the LNA–VGA chain. The thermal noise level for both is approximately 14 nV/$\sqrt{\text{Hz}}$; however, flicker noise dominates the entire frequency range of interest. The 1/f noise corner is approximately 6 kHz, very close to the implemented bandwidth. Although the complementary topology reduces the input-referred thermal noise voltage by half, the flicker noise is still directly incurred at the input. Although we used large area input transistors, the flicker noise still has a significant effect on the overall noise performance. The integrated noise from 0.1 Hz to 25.6 kHz is 1.9 μV_{rms} for LNA alone, and 1.8 μV_{rms} for LNA–VGA chain. Notice that the addition of VGA does not increase the overall noise performance because its noise is reduced by the gain of the first-stage LNA.

5.3 In-Vivo Testing

We used the closed-loop complementary-input low-noise amplifier (LNA2) in an in-vivo neural-recording experiment to verify compatibility with the high source impedance of a neural electrode. We first recorded from traditional rack-mounted instrumentation to identify active spiking cells and then began recording from our proposed circuit.

Because the variable-gain amplifier provided a high-pass corner further attenuating any remaining low-frequency interference, no significant interference was

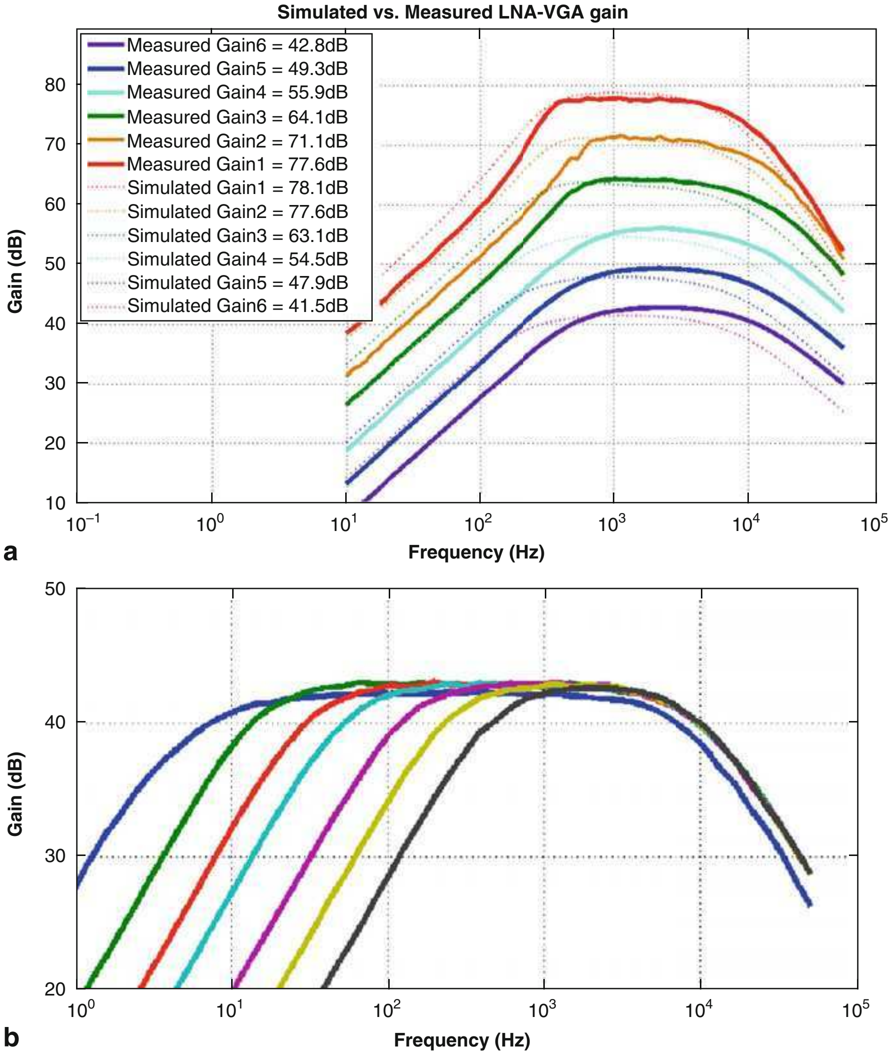


Fig. 5.6 **a** Measured and simulated VGA gain settings with high-pass corners adjusted accordingly; **b** Measured and simulated VGA high-pass corner for the gain setting of 3 dB

observed during the experiment. Our chip produced no noticeable loss of signal to noise ratio (SNR) compared to results obtained using a commercially available rack-mount bioamplifier system. Figure 5.8b shows sorted spikes recorded through our prototype amplifiers. We conclude from these results that we can achieve extremely high fidelity neural recording from a 1 V supply with less than 15 μW power consumption per channel.

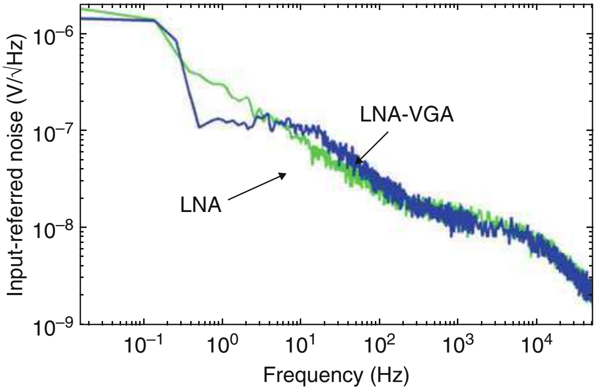


Fig. 5.7 Measured input-referred voltage noise spectrum for LNA and LNA–VGA chain

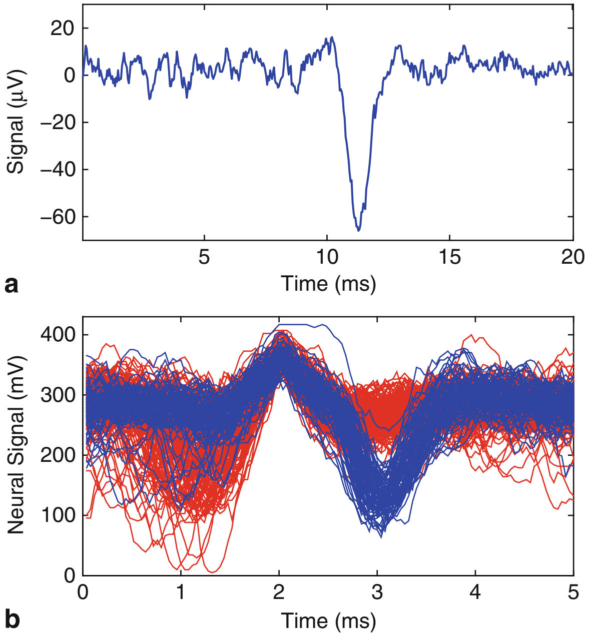


Fig. 5.8 LNA2 tested in vivo in rat motor cortex. **a** Recorded rat spike; **b** Two classes of spikes sorted by post-processing programs

References

[1] Denison T, Consoer K, Santa W, Avestruz A, Cooley J, Kelly A (2009) A 2 µW 100 nV/rtHz chopper-stabilized instrumentation amplifier for chronic measurement of neural field potentials. IEEE J Solid-State Circuits 42(12):2934–2945

[2] Harison R, Charles C (2003) A low-power low-noise CMOS amplifier for neural recording applications. IEEE J Solid-State Circuits 38(6):958–965

[3] Holleman J, Otis B (2007) A sub-microwatt low-noise amplifier for neural recording. In: 29th annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBS 2007), Lyon, pp 3930–3933

[4] Wattanapanitch W, Fee M, Sarpeshkar R (2007) An energy-efficient micropower neural recording amplifier. IEEE Trans Biomed Circuits Syst 1(2):136–147

[5] Wu H, Xu Y (2006) A 1V 2.3 μW biomedical signal acquisition IC. In: Proceedings of the 2006 IEEE international conference on solid-state circuits, Digest of Technical Papers, San Francisco, CA, pp 119–128

Chapter 6
Design and Implementation of Chopper-Stabilized Amplifiers

One observation from the two neural amplifiers described in the previous two chapters is the dominance of flicker noise. Most of the existing power-noise optimization techniques target thermal noise. However, flicker noise is a significant concern for EMG/EEG/ECoG applications, where the bandwidth of interest is much lower (<500 Hz) than that of neural applications (~ 10 kHz). Therefore, we will devote this chapter to discussing techniques to combat flicker noise.

6.1 Chopper-Stabilization Technique

The chopper-stabilization technique is widely used to suppress offsets and 1/f noise. It can be used in applications such as biomedical measurements and human health monitoring. When the signals of interest fall below a few hundred Hertz, the noise that plagues the circuit design shifts away from the thermal noise to 1/f and popcorn noise in transistors [1]. Excess low-frequency noise can undermine the systems signal-to-noise ratio (SNR) and cause errors to the measurement. As a result, chopper-stabilized amplifiers can be effectively used at the front-end of these low-bandwidth signal acquisition applications.

6.1.1 Open-Loop Operation Principle

During open-loop operation, the input signal V_{in} is up-converted by a CMOS switch modulator to chopper frequency (above the low-frequency noise corner) before entering the amplifier. After amplification, a second modulator downconverts the signal back to baseband while simultaneously upconverting the low-frequency flicker noise/offset to the chopper frequency. A low-pass filter restores the desired signal and suppresses the low-frequency noise/offset at the output.

The open-loop architecture contains several limitations. First, transients at the output of the amplifier caused by the finite bandwidth of the amplifiers result in even harmonics at the chop frequency, which in turn create distortion and sensitivity error.

J. Holleman et al., *Ultra Low-Power Integrated Circuit Design for Wireless Neural Interfaces*,
DOI 10.1007/978-1-4419-6727-5_6, © Springer Science+Business Media, LLC 2011

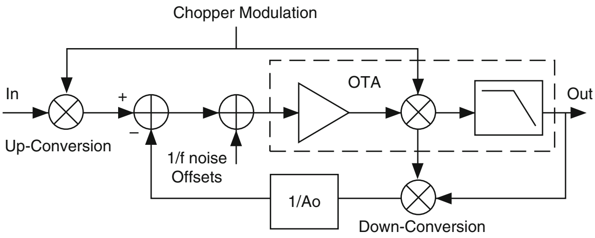


Fig. 6.1 Closed-loop chopper-stabilization technique [1]

Excessive power required to ensure sufficient bandwidth increases power overhead. Secondly, saturation of the amplified offset at the amplifier output limit the first-stage gain, which in turn undermines the input-referred noise from the second stage.

6.1.2 Closed-Loop Operation Principle

Closed-loop feedback techniques can be used to relax the issues mentioned above (Fig. 6.1). A few implementations were published earlier [2, 3], among which [1] has provided the best figure-of-merit to date. In [1], AC feedback paths were employed to ensure all signals entering the amplifier to be well above 1/f noise corner. This technique allowed the use of low-noise on-chip capacitors instead of resistors in the input and feedback signal chains. In addition, he also performed fast modulation within the transconductance stage prior to integration so that the switching dynamics of the chopper is much faster than the chopping frequency. He demonstrated the advantage of the closed-loop technique, in which the gain error and sensitivity are suppressed without further compensation. In addition, he could run the amplifier with low supply overhead to aid in minimizing power without sacrificing noise performance.

6.2 Design of a Chopper-Stabilized Amplifier

We chose a chopper-stabilized topology to suppress 1/f noise and offsets that plague submicron CMOS processes. In order to reduce the signal errors created by amplifier's finite bandwidth, and to relax the headroom constraint on the amplified offsets under low-supply conditions, we adopted a closed-loop feedback technique previously proposed by Denison et al. [1]. We will compare and contrast with [1] in the remainder of this discussion on our prototype chopper-stabilized amplifier.

As shown in Fig. 6.2, a fully-differential closed-loop architecture is used to ensure sufficient linearity and supply rejection. A telescopic-cascode op-amp topology was used. Input transistors are biased in weak inversion to maximize the transconductance efficiency. Dual feedback paths set the mid-band gain of the amplifier through C_{fb}; while another pair biases the amplifier's input node through high-resistance

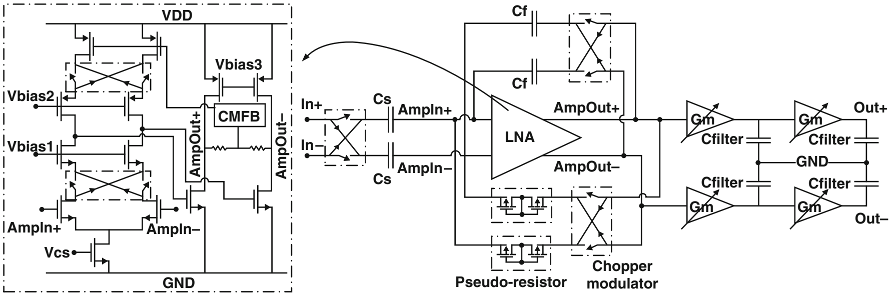


Fig. 6.2 Schematic of the proposed custom chopper-stabilized amplifier

($>$10 GΩ) MOS-bipolar pseudo-resistors. Signal up-conversion occurs at the gate of the input transistors. We introduce a technique that uses chopper switches in both the signal and biasing paths to not only guarantee negative feedback around the amplifier, but also avoid additional input-biasing circuitry as in [1]. We realized the chopper modulator with minimally sized CMOS switches to minimize charge injection. The input capacitance (C_{in}) is 15 pF. When modulated with a 10 kHz chopper clock, the input impedance (1.06 MΩ) is high enough to avoid loading the electrodes for biomedical applications. The ratio of C_{in} and C_{fb} establishes a 40 dB mid-band gain. C_{fb} is sized slightly smaller (140 fF) to take into account the addition of parasitic and switch capacitances. In [1], the total first-stage 100× gain is partitioned into 20× and 5× in order to realize a well-defined high-pass corner with reasonably sized on-chip capacitors. In our implementation, because the precise high-pass corner does not necessarily need to be realized in the first amplifier stage, we achieved 40 dB total gain in one stage. As a result, the input-referred noise from the stages (Gm-C filters) following the chopper-stabilized amplifier could be reduced further by a larger first-stage gain.

The input signal is modulated to the chopper frequency prior to entering the amplifier by a set of chopper switches. Two additional sets of chopper switches are added in the first stage of the amplifier: one set of switches is placed at the drains of the input transistors to demodulate the AC signal down to baseband and modulate the input offsets up to the chopper frequency; another pair is placed at the drains of the PMOS current source to modulate their flicker noise up to high frequency. At the output of the amplifier, the signal returns to baseband while the offsets and flicker noise are modulated up to high frequency and then filtered by the amplifiers 2nd-stage integrator. The 2nd-stage is implemented as common-source topology without tail current source to increase the output swing. The output is then fed back to the summing node at the input of the amplifier after being modulated up to the chopper frequency. In order to avoid large passive devices like the ones in [1], we implemented continuous-time tunable Gm-C filters to attenuate chopper switch ripple at the output of the amplifier. The six bandwidth settings of the Gm-C filters are logarithmically spread between 150 and 400 Hz. The tunability of the Gm-C filters is realized by changing the current in the biasing of the transconductor.

6.3 Hardware Implementation

The prototype was fabricated in 0.13 µm CMOS technology. Everything required for the chopper-stabilized amplifier and filters are on-chip (crystal oscillators, clock generation, etc).

6.3.1 Transfer Function

The measurement result of transfer function (bode plot) matches closely with simulation. Figure 6.3a plots the normalized transfer function of the cases when chopper is on and off (the clock that drives the gates of the chopper switches is turned on and off) for comparison. The mid-band gain for both cases is approximately 38.5 dB. The minute difference with the design (40 dB) is likely caused by parasitic fringe mismatch between the two feedback capacitors. The difference between the two cases lies in the low-frequency high-pass corner. The chopper-off setting demonstrates a high-pass corner at around 0.2 Hz, whereas the chopper-on setting exhibits a passband that extends to DC. The low-pass corner resides around 230 Hz. This corner could be changed by adjusting the current in the post Gm-C filters. Along with the two cascaded Gm-C filters, the overall transfer function exhibits a combined 60-dB roll-off.

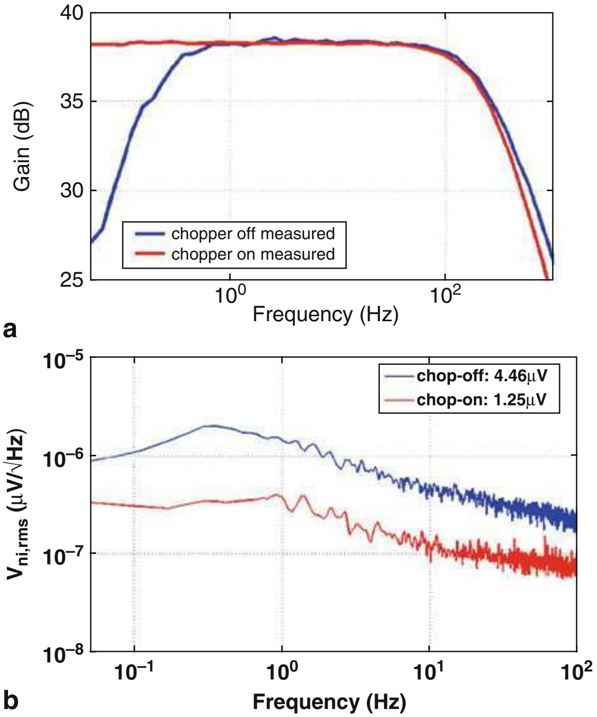


Fig. 6.3 Measured transfer function and noise plot of the chopper-stabilized bio-signal amplifier. **a** Chopper-Amplifier Bode Plot, **b** Chopper-Amplifier Noise Plot

Table 6.1 Comparison of custom chopper amplifier and OPA349

	GBW (kHz)	I_{Amp} (μA)	VDD (V)	$v_{ni,RMS}$(0.1–10 Hz)	CMRR
OPA349	70	1	1.8–5.5	5.7 μV_{rms}	>52
SocWISP chopper amp	20	1.21	1.2	0.4 μV_{rms}	>70

The fully-differential architecture is used to increase the output signal swings and to improve PSRR and CMRR under low-supply condition. CMRR is measured by tying the two differential inputs together and apply a small signal at the common input; PSRR is measured by AC-grounding the input, and apply a small signal at the supply rail. The CMRR and PSRR are not affected by the chopping operation. All results are measured with an HP Dynamic Signal Analyzer (DSA).

6.3.2 Amplifier Noise

The input-referred noise of the amplifier is also recorded as in Fig. 6.3b when chopper switches are turned on and off. The low frequency spot noise is approximately a decade or more lower with chopping enabled. This result corresponds with the operation principle and motivation of chopper-stabilization technique where 1/f flicker noise is reduced by frequency translation and filtering. The 1/f noise corner is found to be approximately 10 Hz, higher than the simulation result (1 Hz). The integrated noise from 0.05 to 100 Hz is measured to be 1.25 μV when the chopper switches are on, compared to 4.46 μV when the chopper switches are off. They come very close to simulated result: 1.1 μV when the chopper switches are on, and 4.75 μV when the switches are off.

Table 6.1 compares the key performance of a commercially-available op-amp and the described custom chopper-stabilized amplifier. Our chopper-stabilized amplifier has much lower noise performance with lower power consumption.

References

[1] Denison T, Consoer K, Santa W, Avestruz A, Cooley J, Kelly A (2009) A 2 μW 100 nV$\sqrt{\text{Hz}}$ chopper-stabilized instrumentation amplifier for chronic measurement of neural field potentials. IEEE J Solid-State Circuit 42(12):2934–2945

[2] Makinwa K, Huijsing J (2001) A wind sensor with an integrated low-offset instrumentation amplifier. Proc ICECS 3:1505–1508

[3] Yazicioglu R, Merken P, Puers R, Van Hoof C (2007) A 60 μW 60 nv$\sqrt{\text{Hz}}$ readout front-end for portable biopotential acquisition systems. IEEE J Solid-State Circuits 42(5):1100–1110

Chapter 7
Spike Detection and Characterization

In neural recording applications focused on action potentials, one of the first signal processing tasks is to distinguish the spikes from noise and interference. In this chapter we will discuss the requirements for a spike detector, review several implementations, and describe in detail a low-power spike detector utilizing the non-linear energy operator and operating in the analog domain. Once a spike is detected, it may be necessary to extract a quantitative description of its shape for subsequent processing stages, such as spike sorting or the detection and rejection of artifacts from electrical stimulation. To this end, the spike detector is combined with feature extraction circuitry, which measures the maximum and minimum of detected spikes, and an ADC, which digitizes the detected values.

In this chapter we describe a circuit to perform spike detection in the analog domain, precluding the need to digitize the entire waveform. After a spike is detected, the maximum and minimum values are digitized with an 8-bit successive approximation ADC. By extracting the most important features of the signal in the analog domain, the power required to digitize the entire waveform is saved. Compared to a simple thresholding scheme, this architecture provides additional information by capturing the maximum and minimum values of the action potentials, which can be used for further processing, including spike sorting or artifact rejection. Additionally, the nonlinear energy operator (NEO), which is used to implement the spike detector, has superior discriminatory ability to a threshold-based detector when the signal is noisy.

7.1 The Spike Detection Task

Spike detection is the task of distinguishing neural action potentials, or spikes, from background noise and interference. An example of spikes with background noise is shown in Fig. 7.1 Noise and interference comes from a variety of sources, both within and outside of the spike frequency range.

Interference from 50/60 Hz line power can be significantly larger than the spikes themselves—often in the tens of millivolts. Because the line power frequency is separated in frequency from most of the energy in the spike signal, its effect can be

J. Holleman et al., *Ultra Low-Power Integrated Circuit Design for Wireless Neural Interfaces*, DOI 10.1007/978-1-4419-6727-5_7, © Springer Science+Business Media, LLC 2011

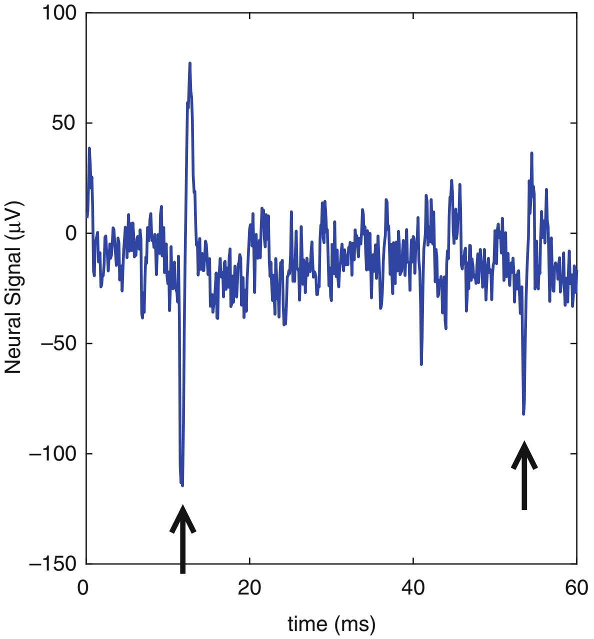


Fig. 7.1 Neural data recorded from a macaque monkey, courtesy of Chet Moritz. The two arrows indicate spikes

greatly reduced by high-pass filtering. However, a number of factors limit the extent to which filtering can improve the situation. Power-constraints in implantable devices will typically preclude the use of high-order filters. Therefore, it may be helpful to place the high-pass corner fairly high when spike detection is the primary goal. For example, a 3rd-order high-pass filter with a cutoff frequency of 250 Hz could be expected to attenuate a 60 Hz tone by 37 dB, meaning that a 50 mV interferer is reduced to an amplitude of about 700 µV, still several times larger than typical spikes. By increasing the cutoff frequency to 1 kHz, the attenuation at 60 Hz is improved to 73 dB, reducing a 50 mV interferer to about 11 µV. Thus the effective ratio of signal to noise and interference may be improved, even though the 1 kHz cutoff also attenuates some of the desired spiking signal. Adding to the difficulty are the harmonics of the line power frequency, which may extend well into the desired band. Harmonics may be due to amplifier non-linearity or to specific electrical equipment, such as flourescent lighting.

Electronic noise from the amplifier and electrode typically consist of components with white and $1/f^{\alpha}$ spectra, with a great deal of the noise falling in the same frequency range as the desired spikes. With appropriate amplifier design, the RMS noise level can be held to 5–10 µV, resulting in minimal degradation of spike detection accuracy.

The aggregate activity of neurons too distant from the electrode to yield distinct spikes [2] contributes yet another interfering signal. Background neural activity is outside the control of the device designer and naturally occupies the same frequency range as desired neural activity, leaving very few options for reducing its effects.

7.2 Spike Detection Techniques

Over the last few decades, neuroscience researchers have developed several algorithms for detecting spikes in noisy signals, and for classifying the spikes from a single waveform according to the neuron that generated them. Most of this work has focused on software implementations which can be used with recorded neural data. Some research has investigated algorithms suitable for implementation in a space- and power-constrained implantable processor.

The simplest and probably most popular spike detection algorithm is a simple amplitude thresholding operation, where a spike is defined as any point in the waveform with a magnitude exceeding the threshold value [11, 22, 23]. Variations on this algorithm use a function to emphasize the difference between spikes and noise. One such function is the non-linear energy operator (NEO), defined as

$$NEO(x) = \dot{x}^2 - \ddot{x}x.$$

The NEO provides a measure of instantaneous energy in the input signal x. For a sinusoid, it is positive and constant, and reduces to the squared product of amplitude and frequency. It has been found to discriminate between spikes and noise better than a simple thresholding detector, particularly when the signal-noise ratio (SNR) is low [17]. Figure 7.2 illustrates how the NEO can emphasize spikes in a neural signal. In [20] Mukhopadhyay found that the NEO provided more accurate spike detection than detectors using prediction error and had lower computational requirements.

Other studies have found a magnitude-thresholding detector with no emphasizing function to perform well. Obeid and Wolf [21] found that a simple threshold detector performed nearly as well as the NEO or a matched filter, and to reform better

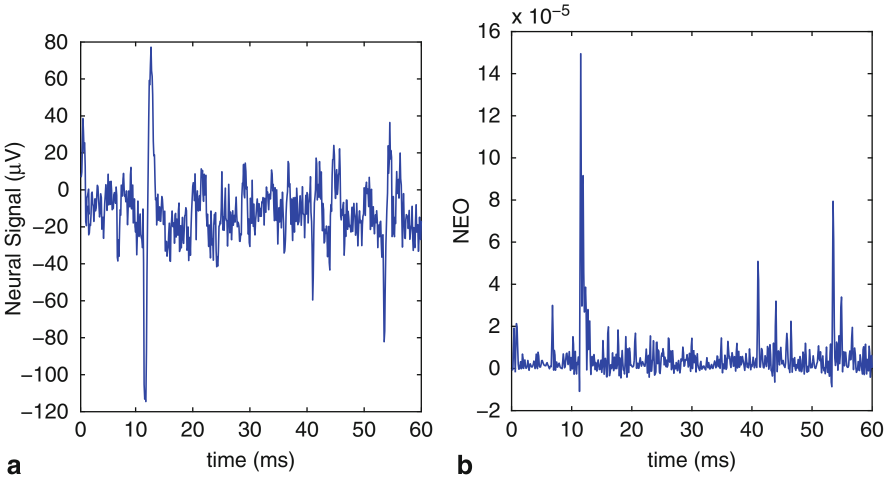


Fig. 7.2 A neural signal (**a**), and the result of applying the NEO to the original signal (**b**)

according to an application-specific cost function accounting for computational costs. In the application studied there, detected spikes were transmitted to a more powerful external computer for verification, so the impact of a false positive was simply wasted transmission power, allowing the cost and benefit of sophisticated detection algorithms to be compared directly. In cases where the entire spike shape is not transmitted, more computational power is justified to reduce the false positive rate.

Whether or not an emphasizing function is used, a threshold-based detector should be capable of adjusting to different signal levels. This can be accomplished by setting the threshold as a multiple of the standard deviation of the noise [9]. A sufficiently high multiple can guarantee a very low false positive rate. For example, assuming white Gaussian noise, threshold levels of 3σ or 5σ would result in probabilities of about 3e-3 or 6e-7, respectively, of a sample being falsely classified as belonging to a spike. Because spikes are only present a small fraction of the time, they do not contribute significantly to the signal standard deviation. Thus the standard deviation of the noise can be safely approximated by that of the signal. Setting the threshold too high can cause valid spikes to be missed.

Spike detection has also been implemented using more sophisticated algorithms. Vogelstein used a support vector machine (SVM) for detection [28] and found the performance to be superior to a magnitude thresholding when the SNR is between 0 and 14 dB. With noisier signals, both techniques failed to provide useful discrimination and with very clean signals both techniques performed well. An SVM-based detector requires training on labeled data, which is undesirable for an autonomous implanted system. SVMs are also computationally expensive, limiting their use in power-constrained implanted devices.

Since action potentials occupy the frequency range roughly between 100 Hz and 5 kHz, adequate digitization would require a neural signal to be sampled at 10 kS/s. With a spiking rate up to about 100/s and a spike width of about 1 ms, around 90% of the digitized samples would not be part of an action potential. These "empty" samples must be digitized and processed using local computer cycles or transmitted via a wireless link for off-chip processing. Either choice results in unnecessary power dissipation. This observation suggests that efficiency may be improved by performing the detection in the analog domain, taking advantage of low-power sub-threshold circuits for the computation. This strategy also reduces the power consumption of the ADC by eliminating the need to digitize the entire waveform.

7.3 Analog and Mixed-Mode Computation

Previous work has shown that for certain applications, analog or mixed-mode signal processing can be more power efficient than fully digital implementations [3, 19]. Specifically, analog circuits have an advantage in power efficiency when the required resolution is low [27]. In [6] Coggins et al. presents a mixed-mode circuit for recognizing cardiac arrhythmias, consuming 200 nW of power. In comparison, a contemporary digital algorithm would have required approximately 375 nW for the analog-digital conversion alone.

Analog circuits are also well suited to spectral analysis. Haddad and Serdijn have demonstrated a continuous time analog implementation of the wavelet transform [8]. The field of continuous-time and discrete-time analog filter design is well established [1, 26]. Harrison et al. reported a circuit for measuring the energy in the 20–40 Hz band of local field potentials in a neural recording. Operating in the analog domain, their circuit consumed 5 nW of power from a 5 V supply.

Two simple circuits which have been proven useful for classification tasks are the bump circuit and the winner-take-all circuit. The bump circuit [7] produces a current which is a function of the similarity between its two input voltages:

$$I_{Out} \propto \mathrm{sech}^2(\kappa \Delta V/2), \tag{7.1}$$

where ΔV is the difference between the two input voltages, and κ is the gate-channel coupling coefficient, a constant for the fabrication process used. The relationship between differential voltage input and current output of the bump circuit is similar to a Gaussian probability density function (PDF), so the current output can be interpreted as a measure of the probability that one of the inputs came from a distribution centered at the other input. The winner-take-all (WTA) circuit [18, 25] takes a number of current inputs, and provides a one-hot encoded binary output indicating the largest input.

A non-volatile analog memory element complementary to analog signal processing techniques can be fabricated using floating gates. In [14], Hsu et al. utilized floating gates to implement an auto-maximizing bump circuit, which continuously adapted a stored value to minimize the difference between the presented inputs and the stored value. The circuit was demonstrated in a simple clustering task. In [16] and [15] the theory of bump circuits for competitive learning is further developed and applied to the task of adaptive vector quantization of handwritten digits.

In [5], Chakrabartty and Cauwenberghs present a pattern classification circuit using floating-gate circuits. During training, a set of templates are learned and stored in floating-gate memories with the chip and a supervising computer in a feedback loop to compensate for circuit imperfections such as mismatch. The circuit classifies 14-dimensional inputs into one of 24 classes and consumes 840 nW of power from a 4 V supply.

In [12], a low-power hardware random number generator utilizing floating gates and mixed-mode signal processing for improved randomness was demonstrated. It included adaptive bias cancellation to improve the random bit distribution, and a programmable mixed-signal FIR filter to remove correlated interference.

7.4 System Design

Based on the considerations discussed in the previous sections, the design discussed here carries out the bulk of the computation in the analog domain. The feature extraction circuit [13], shown in Fig. 7.3, comprises a spike detector for distinguishing

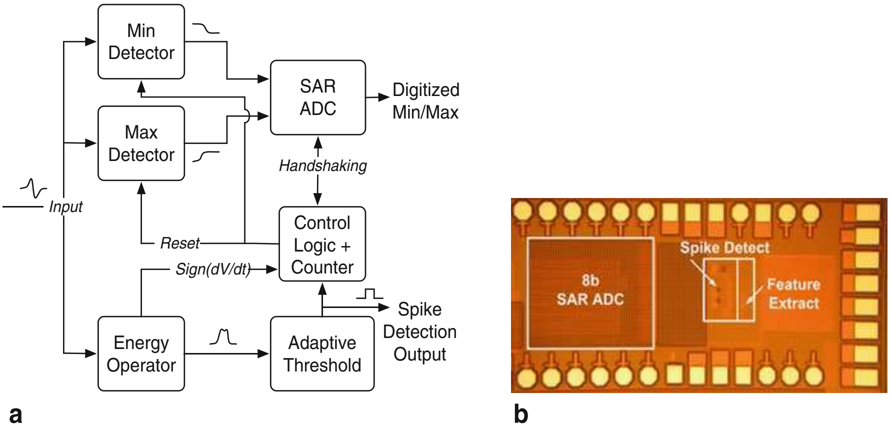


Fig. 7.3 **a** Architecture of the detection and feature extraction circuit. **b** Chip microphotograph

action potentials from noise, positive and negative peak detectors to characterize the detected spike, and a successive approximation register analog-digital converter (SAR ADC) to digitize the spike maximum and minimum. The spike detector is the first component in the signal chain. When a spike is detected, a counter is triggered to provide a delay equal to twice the width of the spike. The delay ensures that the maximum and minimum occur and are captured before the ADC is triggered. After the delay has elapsed, the “Ready” signal is asserted, which causes the ADC to digitize the captured minimum and maximum values. The digitized values are then read through a serial interface. After both conversions are complete, the ADC asserts the “Done” signal, which triggers a reset of the peak detectors and control logic, preparing the system for the next spike detection.

7.4.1 Spike Detector

A schematic of the spike detector is shown in Fig. 7.4. An analog implementation of the nonlinear energy operator (NEO) provides a differential output current which indicates the amount of activity in the input signal.

The two differentiations are performed by gm-C differentiators. The multiplications are performed by Gilbert multipliers. The differential current outputs are connected to perform the subtraction. The multiplier inputs are differential, with the positive inputs taken from the single-ended outputs of the differentiators. The DC levels of the positive multiplier inputs are computed by low-pass filters (not shown) using a pseudo-resistor realized from anti-parallel diodes [10] and connected to the negative multiplier inputs. This arrangement, made possible because there is no useful DC information in any of the signals, prevents offsets in the differentiators from corrupting the NEO output.

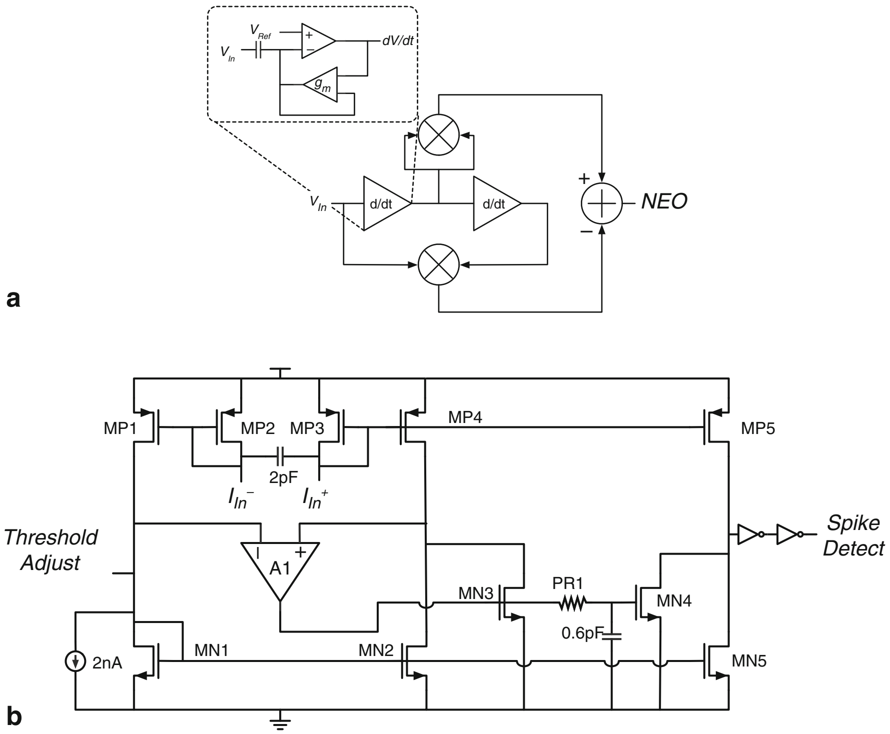


Fig. 7.4 **a** Nonlinear energy operator. **b** Adaptive threshold circuit

An adaptive thresholding circuit converts the NEO output into a binary spike detection signal. Any activity in the input signal, including noise will result in a positive NEO output. In order to minimize false detections, the threshold must be set above the background noise level. The feedback loop formed by A1 and MN3 set $I_{D,MN3}$ equal to the differential NEO input current. This quantity is then low-pass filtered and doubled through the current mirror formed by MN3 and MN4. The low-pass corner frequency is set to around 1–2 Hz by realizing PR1 as a pseudo-resistor formed from anti-parallel diodes. Thus the NEO input required to cause a detection is set at twice the average background activity. The current source in parallel with MN1 ensures that current is flowing through MN3 even when the differential input is zero. A threshold adjustment current can be injected to vary the sensitivity of the spike detector.

7.4.2 Feature Extraction

Positive and negative peak detectors capture the extreme values of the signal. The positive peak detector is shown in Fig. 7.5, and the negative detector is implemented

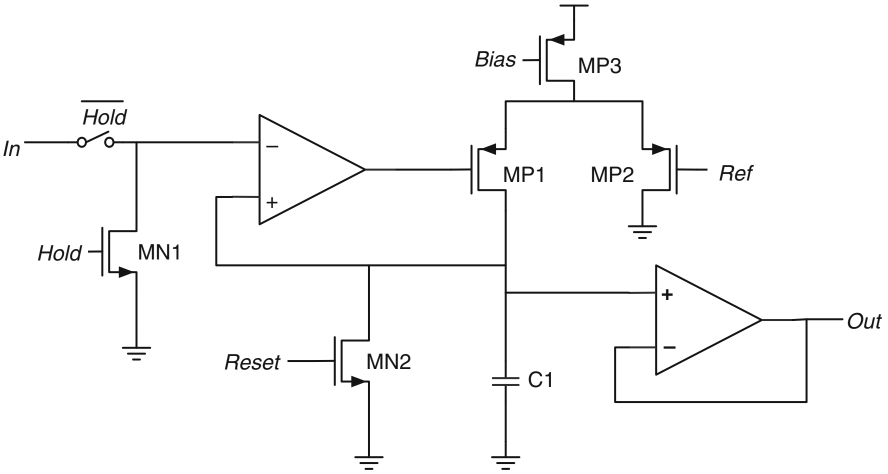


Fig. 7.5 Peak detector

with a similar circuit. The use of a differential pair to charge the storage capacitor allows $V_{SG,MP1}$ to be made less than 0 V, minimizing sub-threshold current in MP1, which could cause the peak detector output to drift to V_{DD} during periods with little activity. When the Ready signal is issued to the ADC, Hold is simultaneously asserted in the peak detector. The peak detector input is forced to 0 V, preventing the output from changing during the analog-digital conversion. After the conversion is complete, Hold is released, and Reset briefly forces the output to 0 V.

A digital counter is used in conjunction with the differentiators from the NEO to measure the width of the spike. The first differentiator has an auxiliary sign output. A change in the sign of the first derivative indicates a minimum or maximum in the input signal. After a spike is detected, the next change in the derivative sign starts the counter. The second change in the sign output causes the counter value to be registered for readout and the counter to count back down to zero. The additional delay allows time for the extreme values of the spike to occur and be sampled by the peak detectors. When the counter returns to 0, the Ready signal is asserted to initiate conversions of the maximum and minimum voltages. The counter is also intended to provide a measurement of the spike width, defined as the time between the maximum and minimum of the spike.

7.4.3 Analog-Digital Converter

The 8-bit analog-to-digital converter (ADC) was designed to operate at 10–100 kS/s. A successive approximation register (SAR) architecture was chosen for the ADC to minimize power consumption [1, 24]. The digital ADC output is read serially from the comparator output. A special sync signal, which is used internally to purge the capacitor array and SAR logic once per conversion, also serves to synchronize the serial output.

7.5 Results

The system was implemented in a 0.13 μm CMOS process. The spike detector and feature extractor occupy a die area of 200 × 220 μm^2. The ADC occupies 295 × 430 μm^2, of which about 85% is consumed by the DAC capacitors.

To test the sensitivity of the spike detector, an artificial neural recording [28] was used. An artificial recording allowed variation of the noise level and spike rate, and provided a reference against which to compare spike detector accuracy. With an actual recording, there is no guaranteed correct reference, since the interpretation of a neural recording is subject to differences in interpretation, even among expert neurophysiologists [29].

Spike detections from the circuit were compared with labels from the generating software to determine the sensitivity and selectivity. For comparison, we also applied a threshold-based software spike detector to the same signal. The software detector indicated a spike whenever the absolute value of the input exceeded a specified threshold. We tested both detectors with several different values for the threshold to build the curves shown in Fig. 7.6. The y-axis shows the false positive rate (FPR), the fraction of detections determined to be false. The x-axis shows the false negative rate (FNR), the fraction of true spikes that were not detected. With a 10 dB SNR, shown in Fig. 7.6a, the threshold-based software detector has good discriminative abilities. Figure 7.6b shows the same curves measured with an SNR of 6 dB. With the noisier signal, the discriminative power of the NEO yields a superior detector at most threshold levels.

To test the accuracy of the digitization, we simultaneously recorded the digital output of the ADC, the timing signals, and the input waveform. We then compared the ADC output to the actual minimum or maximum value that should have been digitized. The comparisons are shown in Fig. 7.7a,b for the maximum and minimum values, respectively. At the end of each pair of conversions, the ADC handshaking signal resets the two peak detectors to allow a new peak to be captured. The true value, plotted on the x-axis, is computed from the recorded input signal as the maximum or minimum value in the time interval between the beginning of a given digitization and the end of the last digitization. The results shown in Fig. 7.7 are for spikes detected when at least 2 ms has occurred since the most recent conversion-reset cycle. The positive and negative peak detectors are reset to 0 V and VDD, respectively, so they become slew-rate limited immediately after a reset signal, causing inaccurate values to be digitized when one spike occurs very shortly after another. Fortunately, neurons have a refractory period of about 3 ms following a spike during which they are unable to generate another action potential, so this is only a significant limitation when multiple cells are being observed on the same channel.

Fig. 7.6 False positive rate (FPR) versus false negative rate (FNR) for the threshold-based software spike detector and the proposed analog NEO detector. **a** SNR = 10 dB. **b** SNR = 6 dB.

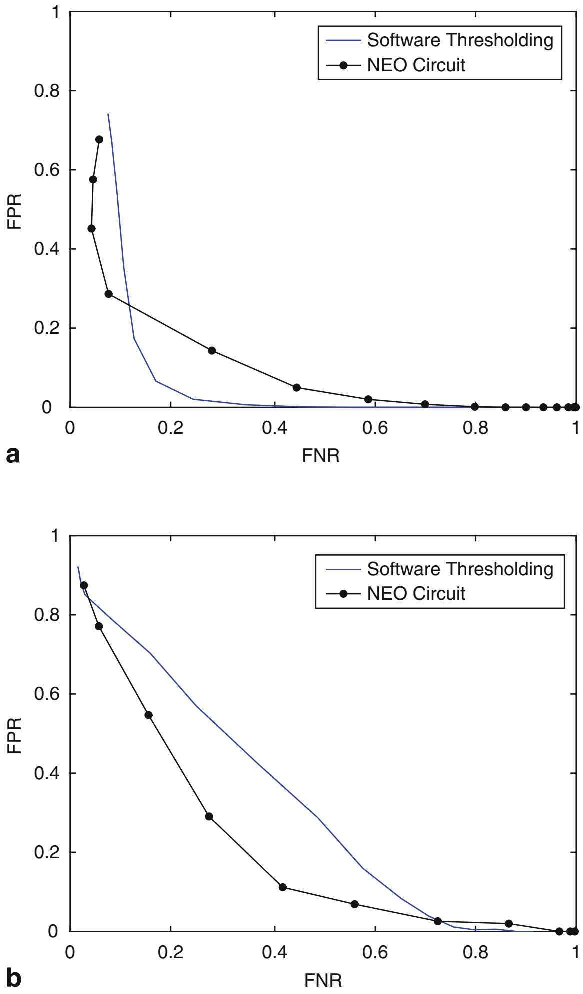
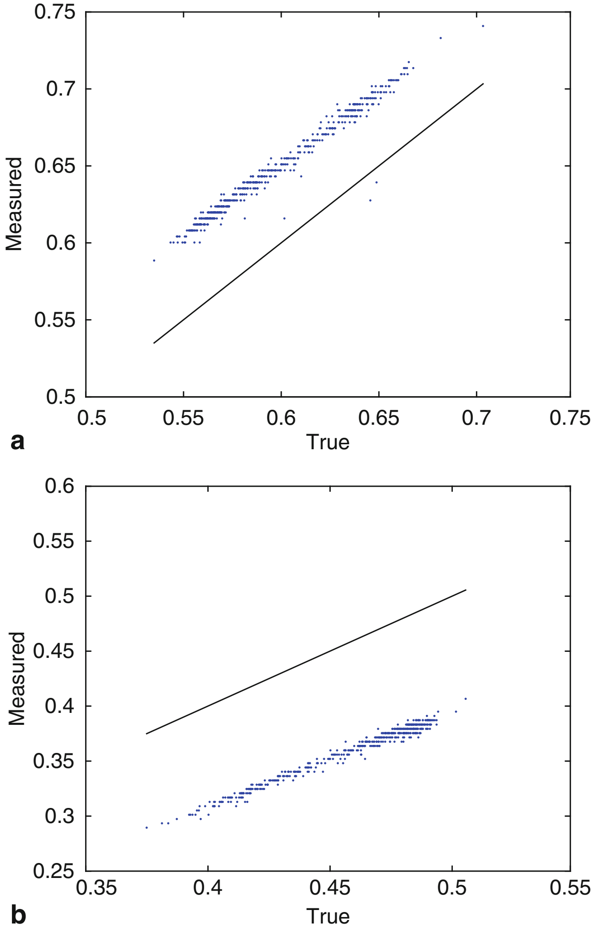


Figure 7.8 demonstrates operation of the complete system. The captured min/max can be seen to track the spike signal. The sign (dV/dt) signal marks the time when the two extreme values occur. The ready signal is asserted when the extreme values are acquired and a delay has passed and reset after the values have been digitized. On reset of the ready signal, the minimum and maximum values are reset.

A performance summary of the spike detection and characterization circuit is shown in Table 7.1. By performing all of the computation using sub-threshold analog circuits, the total power consumption is kept to below 1 µW. Previous work which has implemented similar processing entirely in the digital domain [4], consumed approximately 1 µW/channel to perform spike detection and calculate the maximum, minimum, and width of detected spikes, in addition to the power required for the

Fig. 7.7 Accuracy for capture and digitization of the spike maximum and minimum values. The diagonal line corresponds to the measured value being equal to the correct value



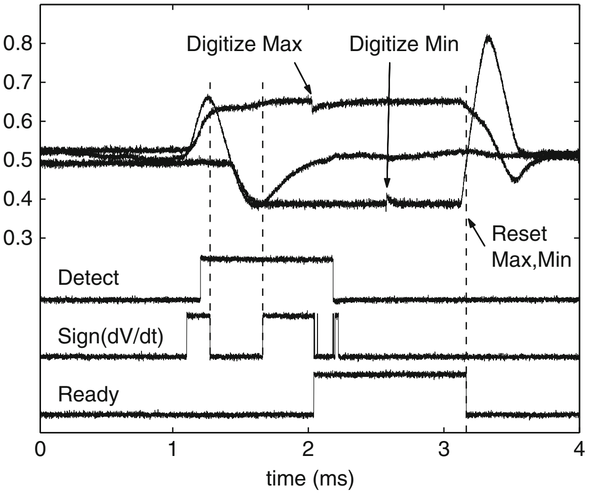


Fig. 7.8 Spike detection timing. Input, V_{Max}, and V_{Min} are shown at the top, followed by the digital timing signals involved in the detection sequence

Table 7.1 Performance summary

Process	0.13 μm CMOS
Area	0.17 mm^2
V_{DD}	1.0 V
Power	0.95 μW

ADC. This strategy can be used to reduce the power consumption required for the combined digitization/processing task. Additionally, because the effective sampling rate of the ADC is so low, a single ADC could be shared amongst a larger number of channels than would be possible with full-waveform digitization, potentially offering a substantial savings in die area.

References

[1] Allen P, Holberg D (2002) CMOS analog circuit design. Oxford University Press, New York

[2] Brown E, Kass R, Mitra P (2004) Multiple neural spike train data analysis: state-of-the-art and future challenges. Nat Neurosci 7(5):456–461

[3] Cauwenberghs G, Edwards R, Deng Y, Genov R, Lemonds D (2002) Neuromorphic processor for real-time biosonar object detection. In: Proceedings of the IEEE international conference on acoustics, speech and signal processing 4:IV-3984–IV-3987

[4] Chae M, Liu W, Yang Z, Chen T, Kim J, Sivaprakasam M, Yuce M (2008) A 128-channel 6 mW wireless neural recording IC with on-the-fly spike worting and UWB transmitter. In: IEEE international conference on solid-state circuits, Digest of Technical Papers, pp 146–147

[5] Chakrabartty S, Cauwenberghs G (2007) Sub-microwatt analog VLSI trainable pattern classifier. IEEE J Solid-State Circuits 42(5):1169–1179

[6] Coggins R, Jabri M, Flower B, Pickard S (1995) A hybrid analog and digital VLSI neural network for intracardiac morphology classification. IEEE J Solid-State Circuits 30(5):542–550

[7] Delbruck T (1991) Bump circuits for computing similarity and dissimilarity of analog voltages. In: Proceedings of international joint conference on neural networks, July 8–12, Seattle, Washington, pp 475–479

[8] Haddad S, Serdijn W (2002) Mapping the wavelet transform onto silicon: the dynamic translinear approach. In: Proceedings of the IEEE international symposium on circuits and systems, IEEE, Scottsdale, Arizona, pp 621–624

[9] Harrison R (2003) A low-power integrated circuit for adaptive detection of action potentials in noisy signals. In: Proceedings of the 25th annual international conference of the IEEE on Engineering in Medicine and Biology Society, vol 4, pp 3325–3328

[10] Harrison R, Charles C (2003) A low-power low-noise CMOS amplifier for neural recording applications. IEEE J Solid-State Circuits 38(6):958–965

[11] Harrison R, Watkins P, Kier R, Lovejoy R, Black D, Greger B, Solzbacher F (2007) A low-power integrated circuit for a wireless 100-electrode neural recording system. IEEE J Solid-State Circuits 42(1):123–133

[12] Holleman J, Bridges S, Otis B, Diorio C (2008) A 3 μW CMOS true random number generator with adaptive floating-gate offset cancellation. IEEE J Solid-State Circuits 43(5):1324–1336

[13] Holleman J, Mishra A, Diorio C, Otis B (2008) A micro-power neural spike detector and feature extractor in 0.13 μm CMOS. In: IEEE custom integrated circuits conference, CICC 2008, San Jose, CA, 21–24 September 2008, pp 333–336

[14] Hsu D, Figueroa M, Diorio C (2001) A silicon primitive for competitive learning. Adv Neural Inf Process Syst 13:713–719

[15] Hsu D, Figueroa M, Diorio C (2002) Competitive learning with floating-gate circuits. IEEE Trans Neural Networks 13(3):732–744

[16] Hsu D, Bridges S, Figueroa M, Diorio C (2003) Adaptive quantization and density estimation in silicon. Adv Neural Inf Process Syst 15:1107–1114

[17] Kim K, Kim S (2000) Neural spike sorting under nearly 0-db signal-to-noise ratio using nonlinear energy operator and artificial neural-network classifier. IEEE Trans Biomed Eng 47(10):1406–1411

[18] Lazzaro J, et al (1988) Winner-take-all networks of o(n) complexity. Department of Computer Science, California Institute of Technology, Technical Report. Morgan Kaufmann, San Francisco, pp 703–711

[19] Lubkin J, Cauwenberghs G (1998) A learning parallel analog-to-digital vector quantizer. J Circuits, Syst, Comput 8(5/6):605–614

[20] Mukhopadhyay S, Ray G (1998) A new interpretation of nonlinear energy operator and its efficacyin spike detection. IEEE Trans Biomed Eng 45(2):180–187

[21] Obeid I, Wolf P (2004) Evaluation of spike-detection algorithms for a brainmachine interface application. IEEE Trans Biomed Eng 51(6):905–911

[22] Quiroga R, Nadasdy Z, Ben-Shaul Y (2004) Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. Neural Comput 16(8):1661–1687

[23] Rao S, Sanchez J, Han S, Principe J (2006) Spike sorting using non-parametric clustering via cauchy-schwartz PDF divergence. In: Proceedings of the 2005 IEEE international conference on acoustics, speech, and signal processing, Institute of Electrical and Electronics Engineers, Piscataway, NJ, vol 5, 14–19 May 2006, pp 881–884

[24] Scott M, Boser B, Pister K (2003) An ultralow-energy ADC for smart dust. IEEE J Solid-State Circuits 38(7):1123–1129

[25] Starzyk J, Fang X (1993) CMOS current mode winner-take-all circuit with both excitatory and inhibitory feedback. Electron Lett 29(10):908–910

[26] Tsividis Y (1994) Integrated continuous-time filter design-an overview. IEEE J Solid-State Circuits 29(3):166–176

[27] Vittoz E (1990) Future of analog in the VLSI environment. IEEE Int Symp Circuits Syst 2:1372–1375

[28] Vogelstein R, Murari K, Thakur P, Diehl C, Chakrabartty S, Cauwenberghs G (2004) Spike sorting with support vector machines. Conf Proc IEEE Eng Med Biol Soc 1:546–549

[29] Wood F, Black M, Vargas-Irwin C, Fellows M, Donoghue J (2004) On the variability of manual spike sorting. IEEE Trans Biomed Eng 51(6):912–918

Chapter 8
Spike Sorting

An electrode in neural tissue can often detect action potentials from multiple neurons. Spike sorting is the task of distinguishing which spikes came from which neurons. It is made feasible by the fact that spikes from a single neuron tend to have a characteristic shape [5].

Early investigations such as [11] focused on finding a minimal set of easily computed features because computational resources were limited. As computation became essentially free on desktop computers, later researchers found that more sophisticated algorithms could achieve more accurate classification. With the power constraints imposed by implantable neural interfaces, it becomes worthwhile to revisit the tradeoff between computational complexity and accuracy.

8.1 Overview

The problem of spike sorting can be roughly broken into three stages: feature extraction, cluster analysis, and classification.

Feature extraction is the process of calculating a small number of parameters to compactly represent a spike, so that the clustering and classification stages can function effectively. Pertinent features can be determined manually at design time and extracted explicitly, or they can be determined automatically using an algorithm like principle components analysis (PCA) [8]. Some researchers have reported good results using only two or three simple features. Zviagintsev et al. developed the Integral Transform [12], which takes advantage of the biphasic shape of most spikes. The integrals of the positive and negative sections of the spike are taken as the two features. They report that sorting based on these two features achieves 98% accuracy on a test data set where sorting based on the first two principle components found by PCA results in 100% accuracy. While the accuracy is somewhat degraded, the computational cost of the Integral Transform is about 2.5% that of PCA. Also, Vibert showed in [11] that a feature vector comprising the positive peak voltage, negative peak voltage, and time between the two peaks is sufficient for accurate sorting.

DOI 10.1007/978-1-4419-6727-5_8, © Springer Science+Business Media, LLC 2011

Clustering is the partitioning of the space spanned by the features into regions such that the points in one region correspond to spikes generated by one neuron. Classification is the process of assigning each spike to one cluster.

Template matching is often used for classification, but requires that the template be known before classification can be performed. In principle it should be possible to devise an automated method for calculating the template, but this adds an additional element of complexity to an autonomous system.

Another category of classifiers is known as time-amplitude window discriminators [1]. In these classifiers an ensemble of spikes is aligned and displayed. The user manually chooses a window in time-amplitude space, corresponding to the waveform having a value in some amplitude range at a given time relative to a threshold-crossing. A spike which passes through a given window is assigned to the corresponding class. Multiple windows can be used to improve discrimination. These discriminators gained early acceptance among neuroscience researchers because of their intuitive operation and because they could be implemented using simple analog circuits, providing real-time classification. Because they require human intervention, time-amplitude discriminators are not suitable for a fully autonomous recording system. Their utility is also limited in systems recording simultaneously from more channels than an individual can monitor. However, they have been used successfully with the window parameters defined at the beginning of the experiment, allowing for autonomous operation thereafter [6].

In recent years, more sophisticated algorithms have been developed for spike sorting. The WaveClus algorithm [7] uses an amplitude threshold for spike detection, and computes a wavelet transform to obtain 64 wavelet coefficients. To choose among the wavelet coefficients, the Kolmogorov–Smirnov test is used to choose the coefficients that have distributions least similar to a gaussian distribution. The rationale is that features useful for sorting will have a multimodal distribution, which will be identified by a high dissimilarity to a gaussian distribution. Clusters are then found using superparamagnetic clustering. Superparamagnetic clustering uses an analogy to statistical mechanics wherein particles are more likely to change state together when they are close together and when the temperature is low. Thus lower temperatures will yield fewer, larger clusters, and at higher temperatures there are many smaller clusters. The temperature can then be chosen based on a minimum cluster size criterion, which for a fixed recording time would lead to a requirement that clusters correspond to a neuron with a minimum firing rate. A more detailed description of superparamagnetic clustering can be found in [7] or [2].

The clustering algorithm developed by Sahani [9], Relaxation Expectation-Maximization, also uses an analogy with statistical mechanics where the number of clusters varies with temperature. Sahani's clustering algorithm is similar to the well-known expectation-maximization algorithm [4] with modifications to automatically choose the number of clusters and to improve convergence. The feature vector in Sahani's spike-sorting algorithm is derived from the waveform using a linear projection similar to that found by PCA.

While the WaveClus and Sahani algorithms both achieve good performance, they are both computationally intensive. They require many iterations over a data set to find the optimal partitioning of feature space. Because of their iterative nature, it is necessary to store a large number of spike waveforms on chip. For a compact low-power implementation, it is desirable to limit the amount of storage and computation required.

8.2 K-Means Clustering Algorithm

K-means is a simple but popular algorithm for finding clusters in data [4]. The "k" refers to the number of clusters, which is chosen before the algorithm begins. Each cluster is represented by its center value, or mean. All of the cluster means are initialized before clustering begins, typically with randomly chosen points from the data to be clustered. K-means is an iterative algorithm, with each iteration consisting of two steps. First, every data point is assigned to the cluster whose mean is closest to it. Then each cluster mean is recalculated as the mean of all of its member points. The process is repeated, and because the means may have shifted, some points may be reassigned to a different cluster. The process is repeated until the means and the assignments do not change.

The standard k-means algorithm requires storage for the entire data set (or some sufficiently large subset) and the computational power to iterate through the data set multiple times. In a system constrained by power consumption and silicon area, it may be desirable to eliminate these requirements. The "on-line" k-means algorithm, illustrated for $k = 2$ and one dimension in Fig. 8.1a, provides an alternative. As with the other versions of k-means, the cluster centers are initialized before clustering begins, but now the data are presented sequentially. As each point is presented, it is assigned to a cluster, and that cluster center is updated to move slightly towards the new data point, according to the equation

$$C_{new} = C_{old}(1 - \lambda) + x\lambda, \tag{8.1}$$

where C is the cluster center, x is the new data point, and λ is a learning rate, which determines how quickly the cluster centers move. With the on-line k-means algorithm, the storage requirement is reduced to that required for the cluster centers themselves. The on-line algorithm is also continuously adaptive, so that if the underlying statistics of the data are gradually changing, the centers can move to track those changes. This also implies that if the centers are moving, then the classification of a given data point may be dependent on when it is received. In the case of spike sorting, it has been shown that the spike shapes do change over time [3, 10], so clustering algorithms should either adapt continuously or periodically re-train.

A similar algorithm is referred to as "fuzzy" k-means in [4], but is equivalent to the Expectation–Maximization (E–M) algorithm with an assumption of identity covariance matrices. In each iteration of this version, data points are assigned with some probability to each of the clusters. That is, for each point, one calculates the probability that the point could have come from a distribution centered at each of the cluster centers, respectively. Then each center is recomputed with an average

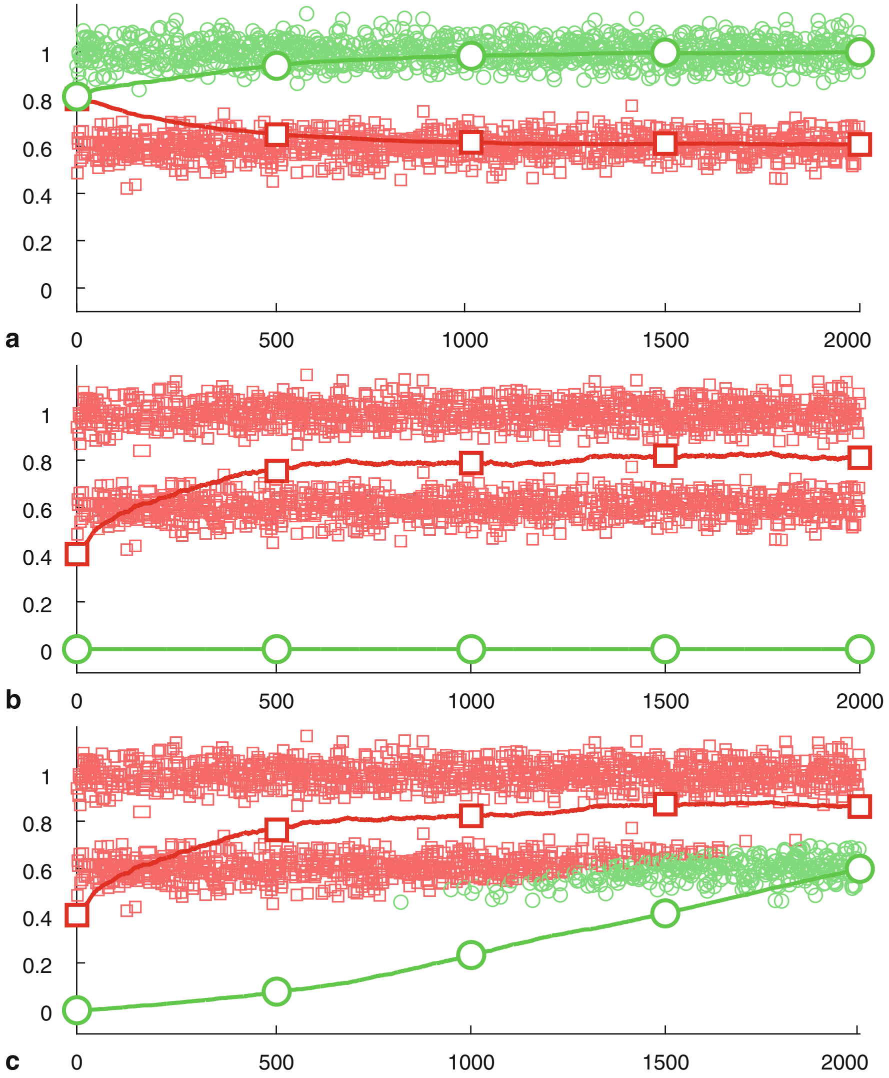


Fig. 8.1 Simulations of clustering the same data with different initial conditions and algorithms. All three simulations use the same data, with two clusters and two centroids in one dimension. The heavy lines with large symbols show the motion of the centroids. The smaller symbols show how each data point is labeled. **a** With both centroids initialized to a point between the two clusters, the k-means algorithm successfully learns the two clusters. **b** With one centroid initialized far from the data, the other centroid incorrectly claims all of the data. **c** With a soft update rule, the E–M algorithm with unity variances is able to recover from the same initial conditions that were problematic for k-means in (**b**)

of all data points weighted by their probability of membership in that class. The process is repeated until the centers and the class assignments change by an amount less than some predetermined threshold. The E–M algorithm is also compatible with on-line implementations. One advantage of this algorithm is that convergence is less dependent on the initial conditions than in the standard k-means algorithm. Consider the case depicted in Fig. 8.1b, with $k = 2$. Center A is initialized far from any of the actual data points, and center B is initialized in the midst of all the data. In the hard k-means algorithm, none of the points will be assigned to A, so A will never be updated. Thus it will sit in its initial position while B will converge to the center of the complete data set. In the simplified E–M algorithm, illustrated in Fig. 8.1c, center A will be assigned all of the points with some small probability, and will gradually migrate towards the data.

One difficulty of the k-means algorithm is that the number of clusters must be determined before the clustering process begins. In some cases, this may require that the number of clusters be determined before the data is even seen. Fortunately, reasonable results can often be attained by simply setting k to the upper limit of the number of clusters that can be reasonably expected. This will result in one underlying spike shape being given multiple labels, but for many applications, this can be addressed in later processing stages.

8.3 Hardware Considerations for Analog On-Line Clustering

Analog circuits can implement a wide variety of computations with very low power dissipation, but typically introduce non-idealities. Such non-idealities may come from device offsets, noise, or the saturating behavior that is an inherent result of a finite supply voltage. Additionally, complexity and power dissipation considerations may encourage approximations to be made in an analog implementation relative to the canonical form of an algorithm, typically intended for software implementation.

This section examines the effect of some approximations used and non-idealities expected in the implementation of an analog clustering circuit.

8.3.1 On-Line Median Learning

In the on-line k-means algorithm, centroids are moved with an update that is proportional to the distance between the new input point and the centroid's current location. Applying a proportional update adds complexity and could be sensitive to offsets. One alternative is to apply a fixed-magnitude update, where the sign of the update is determined to move a centroid towards the new input. In this section, it is shown that in one dimension, a fixed-magnitude update will cause a centroid to converge to the median of the distribution of the input data.

Assume that a sequence of i.i.d. data $x_i \in \mathbb{R}$ for $i = 1, 2, \ldots$ is drawn from an arbitrary probability distribution $P_X(x)$. A quantity c is initialized to an arbitrary value $c_0 \in \mathbb{R}$ and then updated as each datum x_i is received according to the formula

$$c_{i+1} = c_i + \lambda \operatorname{sign}(x_i - c_i), \tag{8.2}$$

where λ is the learning rate.

Assuming a continuous distribution, the probability that $X_i = c_i$ can be neglected, so the expected value of the update is given by

$$E\{c_{i+1} - c_i\} = \lambda(P(X_i > c_i) - P(X_i < c_i)).$$

In the steady state, $E\{c_{i+1} - c_i\} = 0$, so

$$P(X_i > c_i) = P(X_i < c_i). \tag{8.3}$$

Next, recall that the median of a probability distribution is defined as

$$M_X = \operatorname*{argmin}_m \int_{-\infty}^{\infty} P_X(x)|m - x|dx. \tag{8.4}$$

Then we can find M_X by setting the derivative to 0:

$$\begin{aligned}
\frac{\partial}{\partial m}\left(\int_{-\infty}^{\infty} P_X(x)|x - m|\right)dx &= 0\\
\int_{-\infty}^{\infty} P_X(x)\frac{\partial}{\partial m}(|x - m|)dx &= 0\\
-\int_{-\infty}^{M_X} P_X(x)dx + \int_{M_X}^{\infty} P_X(x)dx &= 0\\
P(X > M_X) = P(X < M_X). & \qquad (8.5)
\end{aligned}$$

Since $P_X(x) \geq 0 \ \forall \ x$, the two terms in (8.5) will be monotonically non-increasing and monotonically non-decreasing with respect to x, respectively. Therefore the set of values of M_X that satisfy (8.5) will either be one unique value or a contiguous interval (in the case where $P_X = 0$ over an interval surrounding the median). In either case, any value for c that satisfies the steady-state equation (8.3) also minimizes the total distance cost function in (8.4).

Let $Q_X^N(x)$ be the type of the realized sequence after N draws. The centroid is updated based on the sign of the difference between the input x and the centroid c, so the relevant information in $Q_X^N(x)$ and $P_X(x)$ is contained in the binomial distribution $P_Y(y)$ and the type of its realization $Q_Y^N(y)$, where $y = \operatorname{sign}(x - M_X)$ indicates the direction of the update. By the law of large numbers, as $N \to \infty$, the sample mean $\bar{y}$ of the binomial distribution will asymptotically approach the expected value $E[Y]$. If samples are drawn from below and above M_X with equal likelihood, as is required by (8.5), then $E[Y] = 0$, indicating that as $N \to \infty$, $\bar{y} \to 0$, so the centroid will converge to its steady state where $c = M_X$.

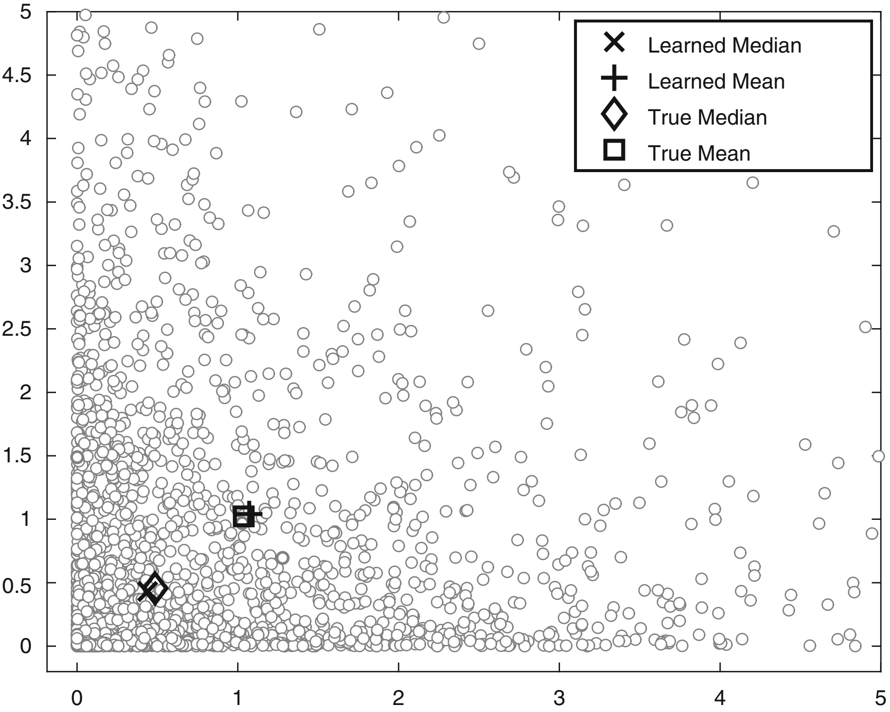


Fig. 8.2 A matlab simulation shows that a learning rule with a fixed update converges to the median, while a learning rule with a proportional update converges to the mean

In practice, c will fluctuate about the median because it is updated at every step. With a sufficiently small learning rate λ, the fluctuations can be made arbitrarily small, at the expense of a longer convergence time.

Figure 8.2 shows simulation results of a single centroid learning a single cluster to compare the results of a fixed-update rule (8.2) and the proportional update rule used in the standard online k-means algorithm (8.1). The data are taken from a two-dimensional Gaussian distribution ($\mu = 0, \Sigma = I$) and squared, so that the mean differs from the median. The diamond and square mark the median and mean calculated from the data. The fixed-update rule converged to the point marked by the '$\times$', which is very close to the calculated median. The proportional-update rule converged to the '$+$', approximating the mean.

8.3.2 Non-Ideal Computational Elements

When an algorithm is implemented using analog circuits, there will be discrepancies between the ideal computation and the actual computation realized. These non-idealities may come from sources including mismatch between elements, noise, asymmetry or non-linearity in the elements' electrical properties.

8.3.3 *Asymmetric Updates*

If the updates are asymmetric, the centroid will not converge to the center of the distribution. If increments are larger than decrements, the centroid will be pushed upwards as long as the number of increments and decrements are equal. As the centroid moves upwards, the proportion of increments will decrease because there will be fewer data points above the centroid. The centroid location will stabilize when the sum of increments and the sum of decrements are balanced. Figure 8.3 shows the effect of asymmetric updates in two dimensions for one centroid adapting to a zero-mean unity-variance Gaussian distribution.

The dependence of the steady-state centroid location on update asymmetry can be determined by noting that if the updates sum to zero, the product of the probability and the magnitude of an increment must be equal to that of a decrement.

$$\lambda_u P(X > c) = \lambda_d P(X < c),$$

where c is the centroid location (in one dimension) and X is a sample from the distribution being learned.

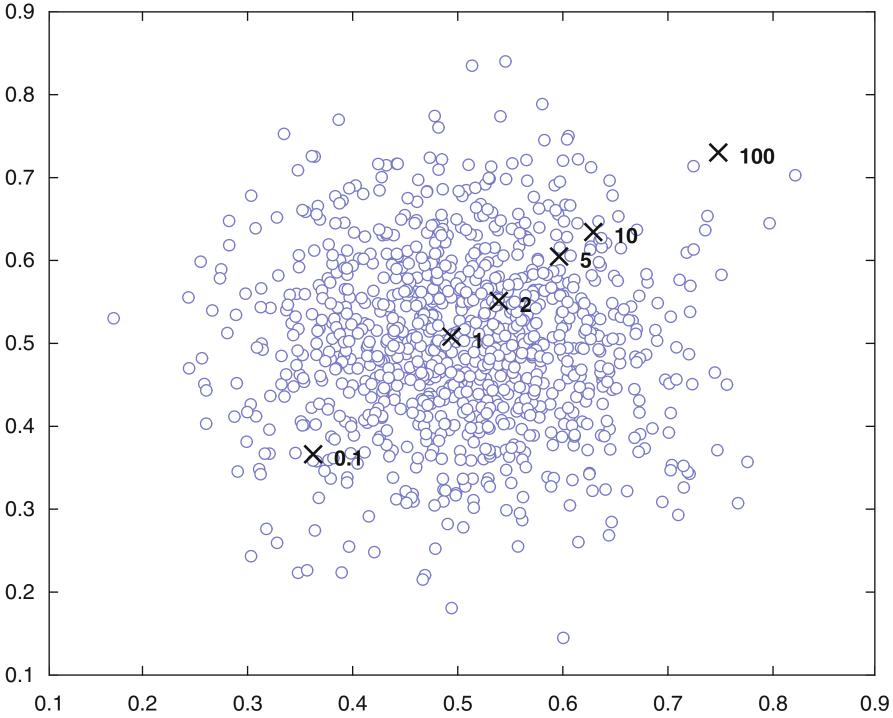


Fig. 8.3 As the ratio between up and down adaptation changes, the learned median moves across the dataset. Here the converged value for 1 centroid is shown for 6 up/down ratios from 1/10 to 100

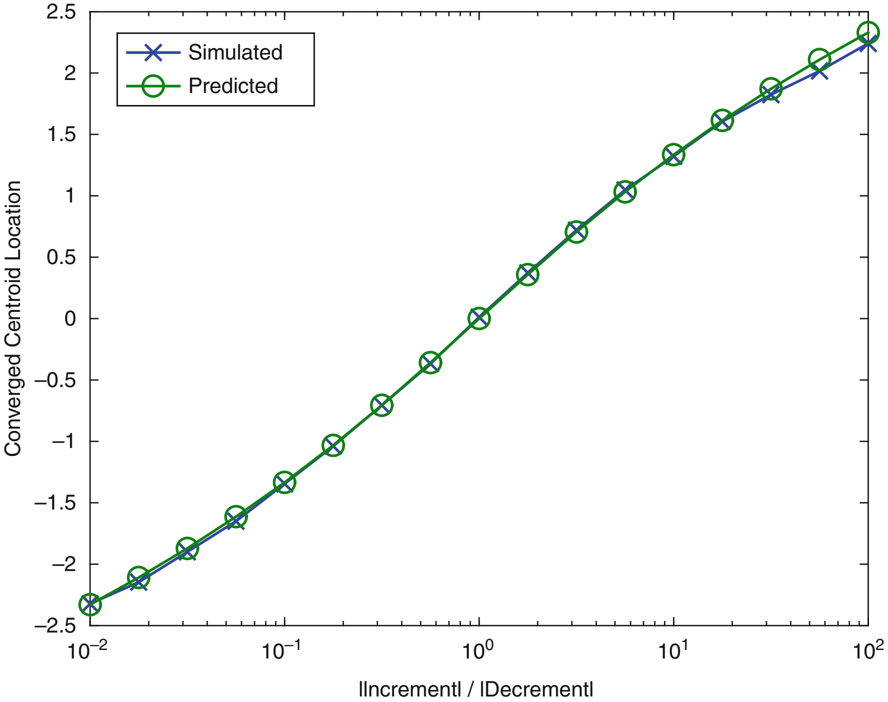


Fig. 8.4 Adaptation was simulated for multiple λ_u/λ_d ratios. The presented data was draws from a 1D Gaussian ($\mu = 0, \sigma = 1$). The value to which the centroid converged is plotted against the ratio and compared to the value predicted by (8.6)

Noting that $P(X > c) = 1 - P(X \leq c)$ and rearranging, we get

$$P(X < c)/(1 - P(X \leq c)) = \lambda_u/\lambda_d.$$

Assuming that X is distributed continuously, so that $P(X = c) = 0$, then the two probability terms on the left correspond to the cumulative distribution function (CDF) of X $F_X(c)$. Define α to be the ratio of increment rate to decrement rate λ_u/λ_d, so

$$F_X(c)/(1 - F_X(c)) = \alpha$$
$$F_X(c) = \frac{\alpha}{1+\alpha}.$$

If the distribution is Gaussian, the CDF can be expressed in terms of the complementary error function erfc $(\cdot)$, yielding

$$\frac{1}{2}\text{erfc}\left(\frac{\mu - \text{c}}{\sigma\sqrt{2}}\right) = \frac{\alpha}{\alpha + 1}$$
$$\frac{\mu - c}{\sigma\sqrt{2}} = \text{erfc}^{-1}\left(\frac{2\alpha}{\alpha + 1}\right).$$

Solving for c shows that the offset of the centroid's steady-state value c relative to the actual center of the distribution can be expressed in terms of the standard deviation σ and the adaptation rate ratio α:

$$c = \mu - \sigma\sqrt{2}\text{erfc}^{-1}\left(\frac{2\alpha}{\alpha+1}\right). \tag{8.6}$$

To verify this relationship, Fig. 8.4 shows the results of several simulations, with the adaptation rate ratios varied across four decades. The converged value of the centroid is compared to the value predicted by (8.6).

Our review of spike sorting algorithms reveals that k-medians is a promising choice for an analog implementation. Our analysis shows that the increment/decrement values must be well matched to achieve an acceptable error. Techniques for achieving this will be described in the next chapter.

References

[1] Bak M, Schmidt E (1977) An improved time-amplitude window discriminator. IEEE Trans Biomed Eng 24(5):486–489

[2] Blatt M, Wiseman S, Domany E (1996) Superparamagnetic clustering of data. Phys Rev Lett 76(18):3251–3254

[3] Brown E, Kass R, Mitra P (2004) Multiple neural spike train data analysis: state-of-the-art and future challenges. Nat Neurosci 7(5):456–461

[4] Duda R, Hart P, Stork D (2000) Pattern classification. Wiley-Interscience, New York

[5] Lewicki M (1998) A review of methods for spike sorting: the detection and classification of neural action potentials. Network: Comput Neural Syst 9:R53–R78

[6] Mavoori J, Jackson A, Diorio C, Fetz E (2005) An autonomous implantable computer for neural recording and stimulation in unrestrained primates. J Neurosci Methods 148(1):71–77

[7] Quiroga R, Nadasdy Z, Ben-Shaul Y (2004) Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. Neural Comput 16(8):1661–1687

[8] Rao S, Sanchez J, Han S, Principe J (2006) Spike sorting using non-parametric clustering via cauchy-schwartz PDF divergence. In: Proceedings of the 2005 IEEE international conference on acoustics, speech, and signal processing, vol 2. IEEE, Toulouse, France

[9] Sahani M (1999) Latent variable models for neural data analysis. PhD thesis, California Institute of Technology

[10] Santhanam G, Linderman M, Gilja V, Afshar A, Ryu S, Meng T, Shenoy K (2007) Hermesb: a continous neural recording system for freely behaving primates. IEEE Trans Biomed Eng 54(11):2037–2050

[11] Vibert J, Costa J (1979) Spike separation in multiunit records: a multivariate analysis of spike descriptive parameters. Electroencephalogr Clin Neurophysiol 47(2):172–182

[12] Zviagintsev A, Perelman Y, Ginosar R (2005) Low-power architectures for spike sorting. In: Proceedings of the 2nd international IEEE/EMBS conference on neural engineering, Arlington, VA, pp 162–165

Chapter 9
Analog Clustering Circuit

This chapter describes the mapping of a clustering algorithm into analog circuits and the design of the constituent circuit blocks. The experimental characterization of the individual blocks and of the clustering system are described. The clustering algorithm implemented is based on the K-Means algorithm, but differs in that the magnitude of the updates is independent of the input.

9.1 Floating-Gate Memories

Floating-gate memories store a value as charge on an isolated poly-silicon node, and convert the charge to a voltage for read-out. Figure 9.1a shows a simple memory cell in which an op-amp A1 is used to buffer the floating-gate voltage V_{FG} to the output. A high voltage ($\geq$7 V) on the TUN input induces Fowler-Nordheim tunneling [8] across the gate oxide of the tunneling junction TJ1, which is simply a pFET with bulk, source, and drain terminals connected. The tunneling process removes electrons from the floating gate, increasing V_{FG}. A high value for DEC will cause the inverter to drive the drain of M1 to 0 V. The channel current in M1 results in the injection of impact-ionized electrons onto the floating gate, lowering V_{FG}. Because the tunneling and injection processes are both extremely sensitive to voltage, the update rates for this structure may vary by orders of magnitude across the output range. For one-time calibration applications, such as offset compensation in amplifiers or comparators, such update rate variation may be an inconvenience. However, for machine learning applications, the update is part of the on-going computation and such update rate variation can lead to large errors or failure of the algorithm to converge.

Figure 9.1b shows a memory cell that uses feedback to hold the floating gate at a constant voltage, in order to remove the dependence of update rate on output voltage [4]. In this memory the op-amp will drive the output to the voltage necessary to keep V_{FG} equal to V_{Ref}. Updates work similarly to the open-loop memory in Fig. 9.1a, except the update operations are reversed because the floating gate is now connected to the amplifiers inverting terminal, so tunneling and injection now decrement and increment the output voltage, respectively.

J. Holleman et al., *Ultra Low-Power Integrated Circuit Design for Wireless Neural Interfaces*, DOI 10.1007/978-1-4419-6727-5_9, © Springer Science+Business Media, LLC 2011

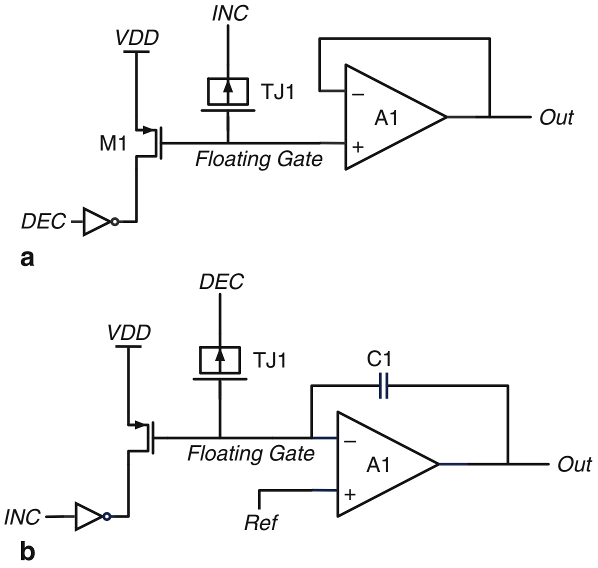


Fig. 9.1 Thick-oxide memory cells. **a** Open-loop memory cell with op-amp buffer. **b** Closed-loop memory cell maintains a constant voltage on the floating gate, stabilizing update rates

9.2 Device Characterization

Until recently tunneling current could be neglected in typical circuits and was therefore not typically modeled by foundry-provided simulation models. As oxides shrink and tunneling current becomes a concern for the larger IC design community, tunneling models are improving, but they may still not have the accuracy needed for reliable circuit design.

In order to guide the design of the floating-gate clustering circuit, floating-gate structures in a 0.13 µm CMOS process were characterized. The memory cell, shown in Fig. 9.2a, has one thin-oxide tunneling device (1 × 0.24 μm^2), a thick-oxide control gate (4 × 20 μm^2), a thick-oxide feedback capacitor (4 × 20 μm^2), and an opamp constructed of thick-oxide devices. The tests were conducted with $V_{DD} = 2$ V and the amp biased with 0.5 µA. The low supply voltage prevented significant current due to

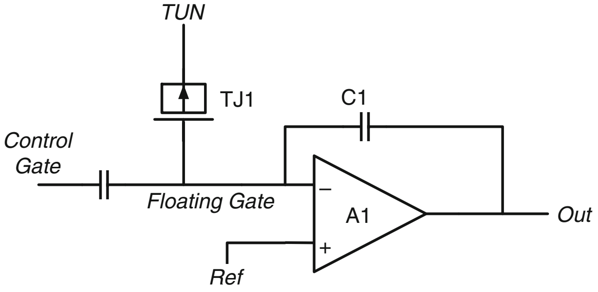


Fig. 9.2 Schematic of the memory cell used for floating-gate characterization

tunneling through the thick-oxide devices or due to hot-electron injection, allowing the tunneling current through the thin-oxide tunneling junction to be measured.

The characterization cell functions similarly to the thick-oxide memory cell in Fig. 9.1b. An ideal op-amp would hold the floating gate at a voltage equal to the reference voltage. Mismatch in the op-amp will shift the floating-gate voltage relative to the reference by the op-amp's input-referred offset voltage. Additionally, finite gain in the op-amp will introduce a slight dependence of the floating-gate voltage on the output voltage. By setting the tunneling voltage (TUN) higher or lower than the reference, current will flow through the tunneling device (TJ1), onto the floating inverting terminal of the op-amp, causing the output voltage to decrease or increase, respectively. Assuming that the only current onto or off of the floating gate is due to tunneling in TJ1, and neglecting effects due to finite op-amp gain, the tunneling current can be estimated as

$$I_{Tun} = \frac{dV_{Out}}{dt} C_1. \tag{9.1}$$

Thus the memory cell allows one to characterize I–V curves involving very small currents ($<10^{-15}$ A) without high-precision instrumentation. It also avoids the need to bring the currents off of the chip, where they would be corrupted by several sources of leakage, interference from AC line power, etc. The measurements presented here were taken by setting TUN to a specific voltage, estimating the gate current from (9.1), and repeating the procedure across a range of voltages.

Here is a quick summary of the interesting results:

- There is an offset in the I–V curves. With apparently 0 V across the tunneling junction, the output voltage drifted by 10–50 μV/s. To hold the output constant required a voltage of about −10 mV across the tunneling device ($V_{\mathrm{Tun}} = V_{\mathrm{Ref}} - 10$ mV). This offset is consistent with the expected amplifier input-referred offset voltage.
- The I–V curves are asymmetric, see Fig. 9.3. This is likely due to the changing density of carriers under the oxide as the device moves through accumulation, depletion and inversion.

In Fig. 9.3 $\frac{dV}{dt}$ is plotted against V_{Tun} for multiple chips. The curves have a significant asymmetry; charge can be tunneled onto the floating gate much more quickly than it can be tunneled off of the floating gate. As discussed in Sect. 8.3, asymmetry can degrade accuracy in learning systems. The asymmetry is most likely due to changes in carrier concentrations on one or both sides of the oxide as the gate-body voltage changes. Also shown is a close-up view of the zero-crossings of the current curves. The reference voltage is 0.8 V, so with an ideal op-amp we would expect a a tunneling voltage of 0.8 V to result in 0 V across the gate oxide, no gate current, and a constant output voltage. The measured zero crossings are at voltages between 5 and 15 mV less than V_{Ref}, which is consistent with the expected magnitude of input-referred offset voltage for the op-amps.

In Fig. 9.4, the time-derivative of the output voltage is used to calculate the gate current. The relationship corresponds to the feedback capacitor being about 7 fF, estimated from simulations.

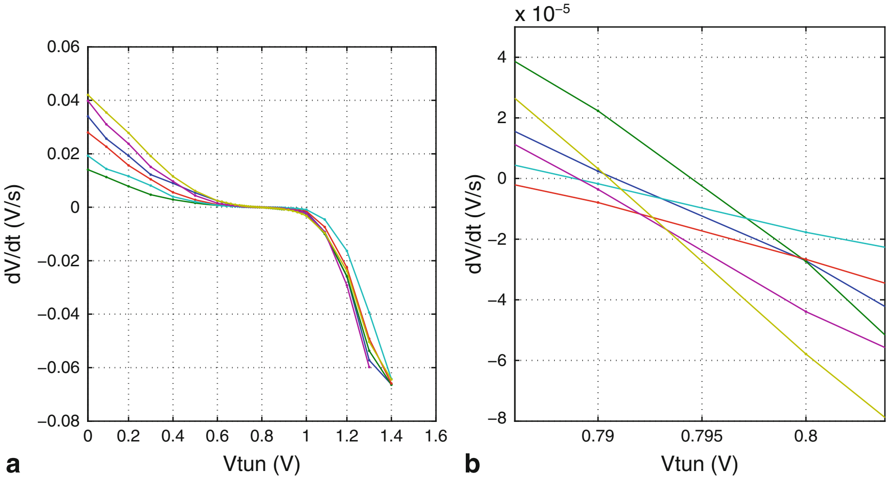


Fig. 9.3 **a** Rate of change of output voltage as a function of the applied tunneling voltage for multiple chips. $V_{Out,Init} = V_{Ref} = 0.8\,V$. **b** A closer view of the same data right around the origin, highlighting the offset voltages

Table 9.1 summarizes the data from Fig. 9.3. V_{Hold} is the interpolated voltage difference across the tunneling junction required to hold the output voltage constant (assumming the opamp to be ideal, so that $V_{FG} = V_{Ref}$). V_{Hold} is the best estimate of the amplifier's offset voltage. Drift (0 V) and Gate Current (0 V) are the rate of change of output voltage, and inferred current onto the floating gate, respectively, when $V_{Tun} = V_{Ref}$.

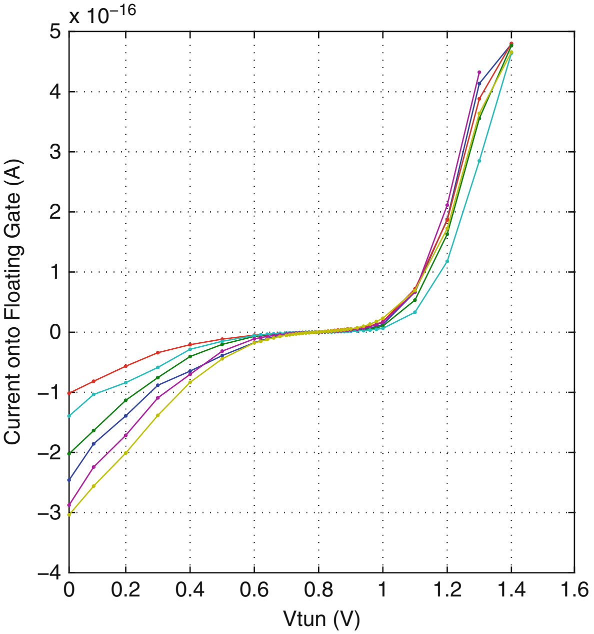


Fig. 9.4 Estimated gate current as a function of V_{Tun} measured across six chips

Table 9.1 Chip variation

Chip	V_{Hold} (mV)	Drift (0 V) (μV/s)	Gate current (0 V)
A	−9.19	−27.13	1.96e–019 A
B	−5.51	−27.33	1.98e–019 A
C	−15.44	−26.64	1.93e–019 A
D	−11.12	−17.75	1.28e–019 A
E	−10.96	−43.92	3.18e–019 A
F	−9.47	−57.90	4.19e–019 A

9.3 Circuit Design

9.3.1 Clustering Circuit

Figure 9.5a shows the clustering circuit, for two clusters and two-dimensional inputs. Each cluster center is stored in one column of the analog memory cells, and each dimension (each feature) is represented by one row. Every cell in the array has a floating-gate memory cell and a difference circuit. The difference circuit is a Gilbert multiplier connected as in Fig. 9.5b to compute the squared difference between two input voltages. When a new input vector is presented, the difference circuits in each cell output a current representing the distance between the stored and presented values. For small differences, the distance current is proportional to the squared difference between the two values. For each cluster center, the current outputs corresponding to each dimension are added to provide a current proportional to the Euclidean distance between the stored vector and the input vector. This distance current in turn provides a measure of the probability that the input vector came from a probability distribution centered at the stored cluster center.

The differential currents from the difference circuits are converted to single-ended current and fed into a loser-take-all (LTA) circuit. The LTA circuit [9] takes multiple current inputs, and provides a one-hot encoded binary output indicating the smallest input. The smallest input current corresponds to the smallest distance and indicates the class to which the input vector is most likely to belong. Each binary output is connected to all of the memory cells in the corresponding cluster center. When the LTA output for a given cluster is high, indicating that it is closest to the presented input, the memory cells adapt towards the input. Thus, each center will converge to the median of all of the inputs that have been assigned to it.

The on-line expectation-maximization (E–M) algorithm has been shown to have convergence properties superior to the K-means algorithm, because in the E–M algorithm every center is updated at every step. The K-means algorithm can be modified to implement E–M with uniform variances by making the updates proportional to probability of class membership.

In order to approximate the behavior of a uniform variances E–M algorithm with minimal circuit complexity, the circuit described here allows for two-level discrete updates. That is, the “winner” (the center to which the input has been assigned) is updated by one amount, and all of the other cells are updated by a much smaller amount. This is conceptually similar to quantizing the class membership probabilities

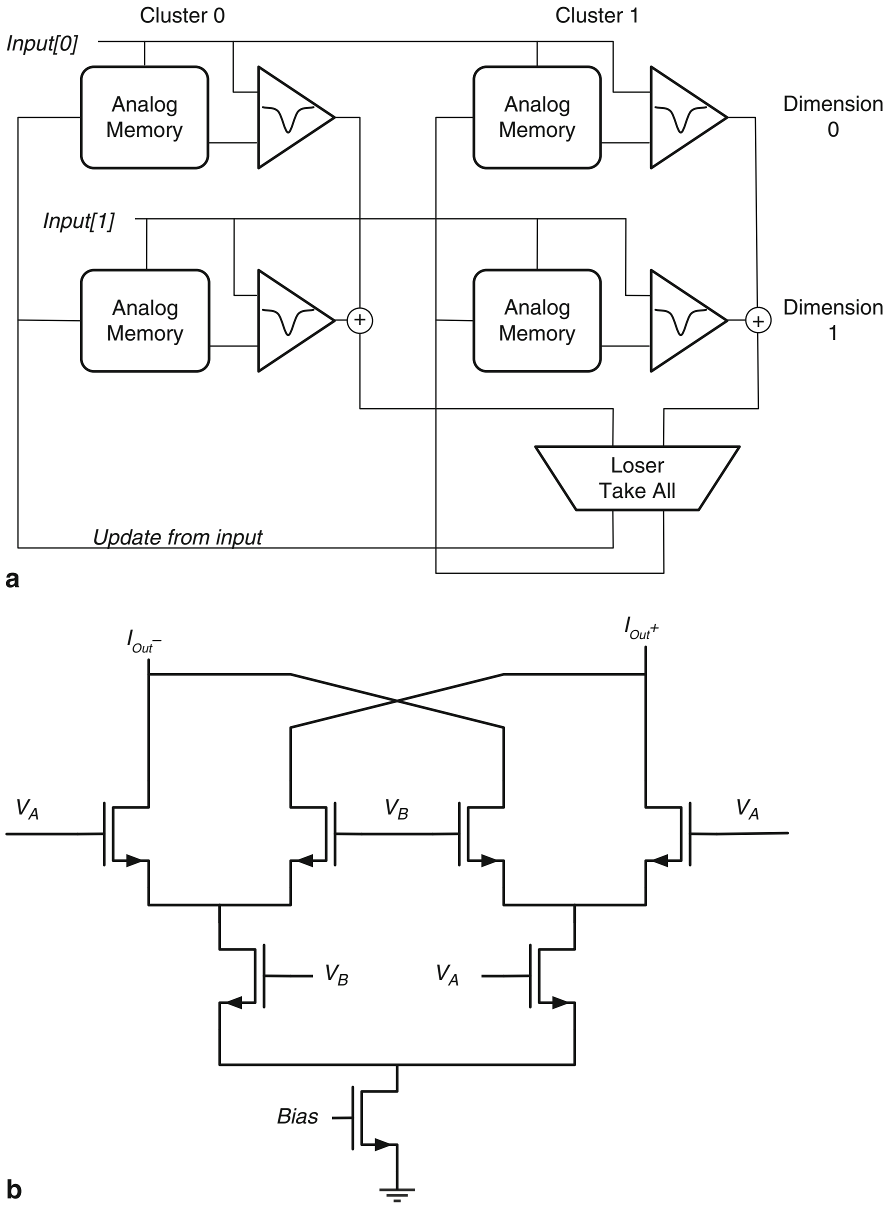


Fig. 9.5 **a** Block diagram of the clustering circuit. The actual clustering circuit implemented has four clusters and three input dimensions. For simplicity, the figure shows two clusters with two dimensions. **b** A Gilbert cell connected to compute the squared difference between two voltage inputs

in the E–M algorithm to two levels. Using the fixed-update memory cell described above, the magnitude of the update is controlled by varying the duration of the update pulse. The update pulses are driven by two external signals, one for the winner and one for the losers.

An alternative topology for a clustering circuit based on floating gate memories [7] uses a bump circuit [3, 6] to compute the one-dimensional (1D) distance between the stored value and the input. The bump circuit generates a current which can be interpreted as the probability that the input voltage is a member of a probability distribution centered at the stored voltage. The disadvantage of this topology is that aggregating the 1D probability into a class membership likelihood requires all of the 1D probabilities to be multiplied. A multiple-input multiplication can be realized with analog circuits, but adds complexity, consumes extra power, and introduces additional offsets. In the topology described here, the 1D squared difference currents can be added by simply connecting the corresponding wires, avoiding the disadvantages of the multiplication circuit.

9.3.2 Floating-Gate Memory Cell

Floating-gate memory cells are used to store the centroid locations. The memory stores a value as charge on an isolated piece of polysilicon, the floating gate. The charge is then converted to an output voltage by an op-amp in a negative feedback configuration.

Figure 9.6 shows two versions of a floating-gate memory cell with different update dynamics. The memory cell is based on the circuit presented in [4], with modifications to allow the use of a thinner tunneling oxide. The tunneling junction TJ1 is simply a PMOS transistor with its source, drain, and well connected together. Voltage between the source/drain/well terminal and the gate terminal causes a current to flow through the gate oxide due to quantum tunneling. Both operate in a hold mode, where the output voltage is held constant, when the Adjust input is low, and in an update mode, where the output is adjusted to be closer to the Target voltage, when the Adjust input is asserted.

Hold mode operation is the same in both cells. The output of amplifier A1 is coupled back to its inverting input through a feedback capacitor. The amplifier drives the output to the voltage necessary (within the limits of supply voltage and output range) to keep the floating-gate and reference voltages equal. This keeps the voltage difference across the tunneling junction equal to the input-referred offset voltage of the amplifier, on the order of 5–10 mV. By minimizing the voltage across the tunneling junction, this structure allows for the use of thin gate oxides (2–3 nm) for tunneling, which can leak even with voltage differences of less than 0.5 V. In contrast, thicker oxides (7–8 nm) experience negligible tunneling current with voltage differences of less than 5 V.

The output voltage is adjusted by varying the charge on the floating gate. This is accomplished by inducing a potential difference across the tunneling junction, which causes a small tunneling current to flow through the oxide. In Fig. 9.6a the target voltage is connected to the feedback capacitor, changing the floating-gate

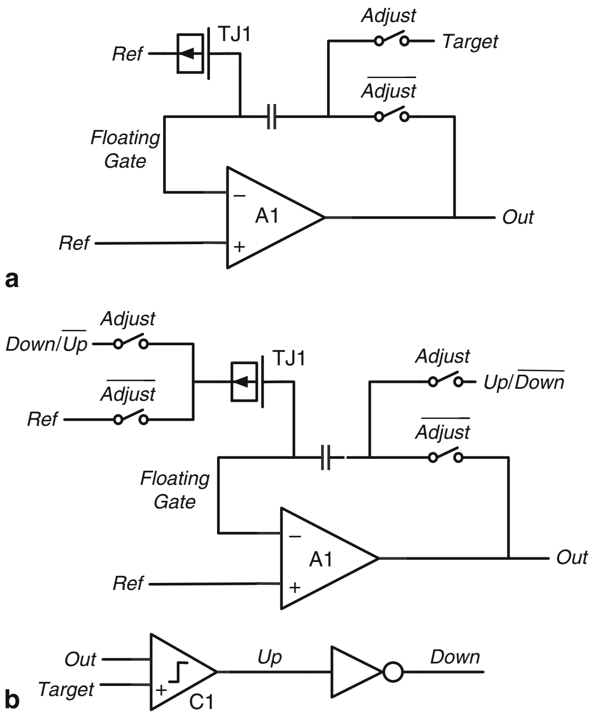

Adjust
TJ1
Ref
Target
$\overline{\text{Adjust}}$
Floating Gate
–
A1
Out
Ref
+
a
Adjust
Down/$\overline{\text{Up}}$
TJ1
Adjust
Up/$\overline{\text{Down}}$
$\overline{\text{Adjust}}$
Ref
$\overline{\text{Adjust}}$
Floating Gate
–
A1
Out
Ref
+
Out
Up
Down
Target
+
C1
b

Fig. 9.6 Two floating-gate memory cells which update towards the Target input when the Adapt signal is asserted. **a** The magnitude of the update is approximately exponential with respect to the difference between the stored value and the Target input. **b** The magnitude of the update is independent of the difference between the stored value and the Target value

voltage. If the target voltage is higher than the current output voltage, the floating gate will be forced up, causing current to flow off of the floating gate through the tunneling junction. When the Adjust signal is deasserted and the feedback loop is re-established, the output will now be forced higher to keep the floating gate at the reference voltage. One benefit of this topology is that the output will asymptotically approach the target voltage. If the Adjust signal is asserted for a sufficiently long time, the floating gate will eventually reach the reference voltage with the target voltage connected to the feedback capacitor. When the output is re-connected to the feedback capacitor, the voltage it needs in order to equalize the reference and floating-gate voltages is equal to the target voltage, plus the product of the input-referred offset of the amplifier and the coupling loss from output to floating gate. The coupling loss is determined by capacitive division between the feedback capacitor and parasitic capacitance on the floating gate node.

The disadvantage of the above memory cell is that as the output approaches the target value, the update rate becomes extremely small, potentially leading to unreasonably long settling times for machine learning algorithms. The memory cell in Fig. 9.6b achieves a larger update rate by applying a larger voltage difference across the tunneling junction. A comparator C1 determines whether the output voltage should be adjusted up or down to move closer to the target voltage. If the target voltage is higher than the current output voltage, the Up signal is equal to the supply voltage V_{DD} and the Down signal is 0 V, so the feedback capacitor is connected to the positive supply, raising the floating-gate voltage. At the same time, the well/drain/source terminal of the tunneling junction is connected to 0 V. Because the floating gate is at a much higher voltage than the opposite terminal of the tunneling junction, current flows off of the floating gate, reducing its voltage, and requiring the output to settle to a higher voltage after the memory is returned to hold mode. The update magnitude is independent of the target voltage for this topology.

9.3.3 Decision Circuit

The classification decision is made based on a vector of distance currents by a loser-take-all (LTA) circuit, shown in Fig. 9.7. The LTA circuit operates as follows: M_1 and M_2 form a current mirror which converts the input currents to the correct polarity and isolates the input current sources (in this case, the Gilbert cells) from voltage changes in the LTA circuit. Initially, the Adapt input is off, disabling the tri-state inverter I1. Transistors $M_{3,4}$ form a self-biased cascode current mirror. A low-power device with a higher threshold voltage is used for M_3 (marked "hvt") so that the same gate voltage can be used to bias M_4 and M_3. A negative feedback loop formed by the $M_{3,4}$ current source and the PMOS source follower M_5 ensures that the sum of $M_{5,i}$ currents across all LTA cells will equal the bias current sourced onto the Common

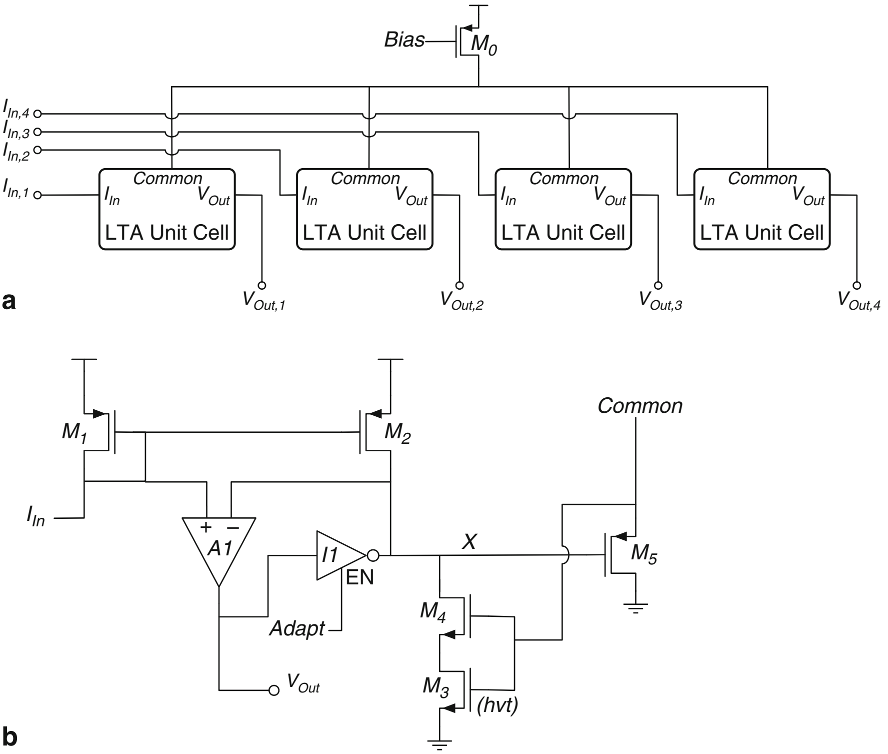


Fig. 9.7 **a** The loser-take-all circuit for finding the smallest distance current. **b** Schematic of the LTA cell, the unit element of the loser-take-all circuit

node by M_0. Because the $M_{3,4}$ current sources in each cell are all controlled by the Common node, they will all sink the same current (assuming they are all operating in the saturation region, and neglecting channel-length modulation). Suppose that all of the input currents are equal and that the I_{M0} is evenly distributed amongst all cells. Then V_{Common} will rise to the voltage required to sink $I_{In} = I_{M1} = I_{M2}$ through $M_{3,4}$. The negative feedback loop formed by $M_{3,4,5}$ will drive V_X to approximately one threshold voltage of M_5 below V_{Common}. The use of a higher threshold for M_3 and the choice of a lower aspect ratio for M_3 (W/L = 3 µm/3 µm) than for M_5 (W/L = 12 µm/1 µm) keep V_X high enough for the $M_{3,4}$ current source to remain operational.

If $I_{In,i}$ decreases, then $V_{X,i}$ will also decrease, bringing V_{Common} down with it. In the other cells, this will cause $V_{X,j \neq i}$ to increase, because $I_{M3,j}$ has decreased, while $I_{M2,j}$ has not changed. Thus cell i with the lowest input current will sink the majority of the bias current I_{M0}, and $V_{X,i}$ will be very low. For the other cells $j \neq i$, $V_{X,j}$ will be high. A1 acts as a comparator to convert V_X to a level compatible with logic inputs.

At each step of the clustering algorithm, when the memory cells adapt, the adaptation mechanism of the memory cells cause their output voltages to be temporarily driven to the supply rails. To prevent these changes from translating to a change in the LTA output, the LTA cells latch their output using a positive feedback loop formed by A1 and I1.

Figure 9.8 shows the operation of the LTA circuit. Initially, $I_{Diff,1}$, the differential input current to cell 1, is the smallest of the input currents, so Winner$_1$ is high while all

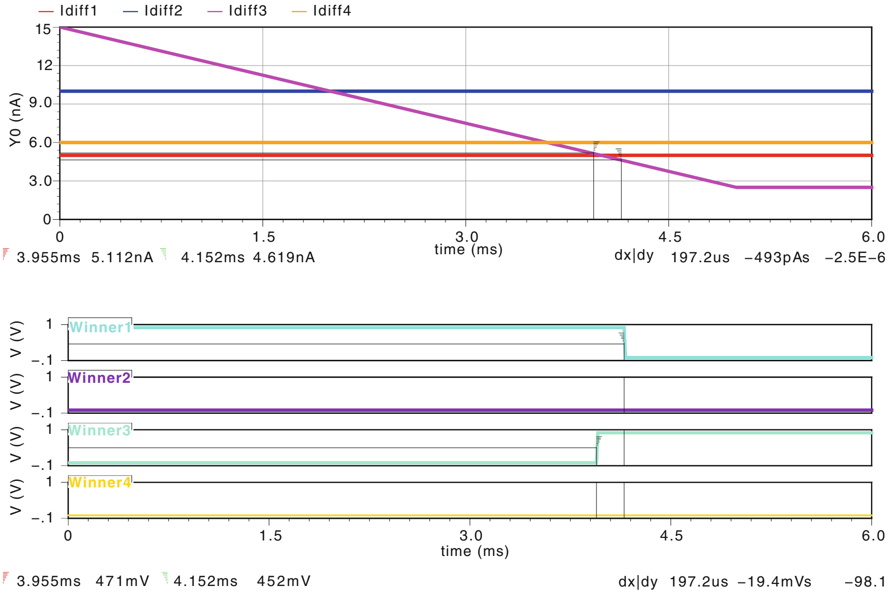


Fig. 9.8 Simulation of the loser-take-all decision circuit. The class 3 input $I_{Diff,3}$ current starts out as the largest and decreases, becoming the smallest at t = 4 ns. When the $I_{Diff,3}$ drops below $I_{Diff,1}$ there is a brief interval, denoted by the tick marks, where the indicator outputs for classes 1 and 3 are both high, corresponding to a current range of about 0.5 nA.

of the other Winner$_X$ outputs are low. $I_{Diff,3}$ decreases, dropping below $I_{Diff,1}$ at 4 ms. Immediately after that, Winner$_3$ becomes high at 3.964 ms, marked with the first cross-hair. For a brief interval, both Winner$_3$ and Winner$_1$ are high, until $I_{Diff,3}$ drops to more than 0.4 nA below $I_{Diff,1}$ and leaves Winner$_3$ as the only high Winner output.

The multiple-winner region is a result of the finite output impedance of the current sources formed by M_2 and $M_{3,4}$. To see this, remember that V_{Common} will be at the voltage required to sink $I_{M2,min}$. If $M_{3,4}$ and M_2 have infinite output impedance, then for any cell j where $I_{M2,j} > I_{M2,min}$, $V_{X,j}$ will be driven to near V_{DD}. Because M_2 and $M_{3,4}$ have finite output impedance, $I_{M2,j}$ may be slightly larger than $I_{M3-4,j}$ without forcing $V_{X,j}$ to V_{DD}.

9.4 Experimental Results

The clustering circuit occupies 0.11 mm^2 in a 0.13 µm CMOS process. The results discussed here were obtained with a supply voltage of 1.5 V, and the circuit biased to consume 3.5 µA.

9.4.1 Update Rates

The update rate of the memory cell depends on the direction of the update, the stored value, the supply voltage, and the reference voltage. In the design proposed here, the reference voltage is accessible and can be used to vary the relative magnitude of increment and decrement magnitudes.

Referring back to Fig. 9.6b, when an increment is initiated, the well (left) terminal of TJ1 is switched to 0 V from V_{Ref}. An increase in V_{Ref} increases the magnitude of the voltage applied across TJ1 as a result of the switching on TJ1's well terminal, increasing the rate of voltage change. Conversely, a decrement results in TJ1's well terminal switching from V_{Ref} to V_{DD}, so an increase in V_{Ref} decreases the magnitude of the voltage applied across TJ1, reducing the decrement rate.

Figure 9.9 shows the increment and decrement rates for three different reference voltages for 1 memory cell. It can be seen that as V_{Ref} increases, the increment rate (dashed lines) increases, and the decrement rate (solid lines) decreases. As described in Chap. 8.3, asymmetry in the update rates leads to an offset in the learned value relative to the actual center of a distribution. The reference voltage can be used to adjust the the ratio of increments and decrements.

The update rate also depends on the value stored in a memory cell. This is a result of the switching on the driven (non-floating, right side) terminal of the feedback capacitor. Before an update is initiated, the voltage of the driven terminal is V_{store}, because it is connected to the amplifier output. When an increment is initiated, that terminal is driven to V_{DD}. As the stored value increases, the step applied to the capacitor, and coupled onto the floating gate decreases in magnitude, causing the increment rate to decrease. For a decrement, the driven terminal of the capacitor is driven down to 0 V, so as the stored voltage increases, the step size and thus the update rate increases.

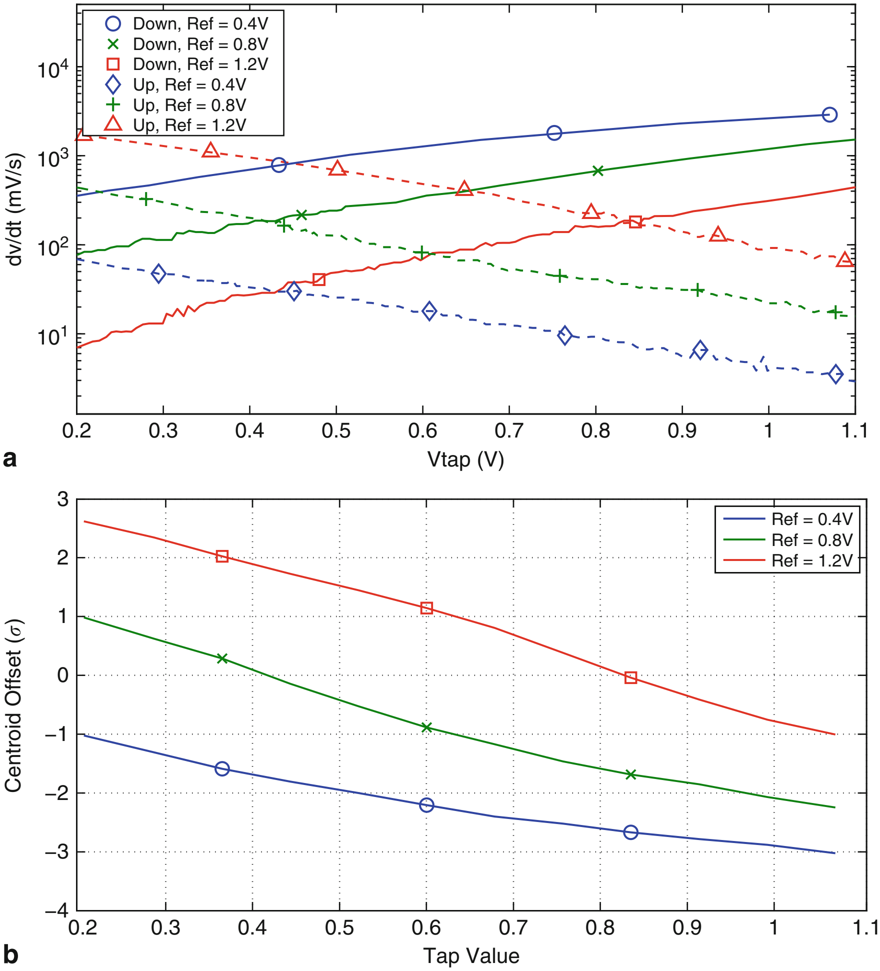


Fig. 9.9 **a** Adaptation rates plotted for three different reference voltages versus stored voltage for programming up (solid line) and down (dashed line). **b** The relative position in a gaussian distribution to which the memory cell can be expected to converge, as a function of stored voltage. The position is expressed as on offset relative to the mean in terms of standard deviations. It is based on the values in (**a**) and the relationship defined in (8.6)

Figure 9.9a shows that for all three reference voltages, the increment rate decreases as the stored voltage increases, while the decrement rate increases. The ratio of the increment and decrement magnitudes can be used to predict the error in a learning application, using (8.6). In Fig. 9.9b, the predicted offset is shown as a function of stored tap value and reference voltage. Because the learned value depends on the inputs, the offset is relative to the center of the distribution, and is proportional to the standard deviation (σ) of the distribution. The tap will be stable when its value in σ relative to the mean is equal to the steady-state value shown in Fig. 9.9b for its value and reference voltage.

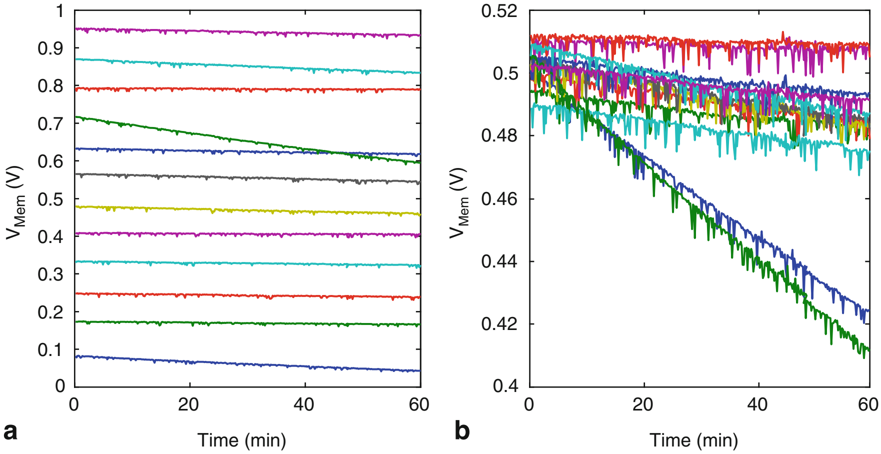


Fig. 9.10 One hour retention of 12 memory cells. **a** Cells initially programmed to voltages ranging from 0.1 to 0.9 V. **b** All cells programmed to approximately 0.5 V. The worst drift measured in these two cases is 124 mV (approximately 2 mV/min) for tap 9, which was initialized to 0.72 V

9.4.2 Memory Cell Retention

In order to test the retention of the analog memory cell, each of the 12 cells on one chip were programmed to voltages evenly spaced from 0.08 to 0.96 V and monitored over the course of 1 h. The results are shown in Fig. 9.10a. The voltage drift varied from −1.7 to −124.0 mV, corresponding to a worst case droop rate of −34 μV/s. In another experiment, the cells were all programmed to approximately 0.5 V, and again monitored over the course of an hour. In this case, the drift ranged −1.7 to −85 mV.

Voltage drift was also measured at 25°C and 75°C to get a rudimentary indication of the temperature dependence of the memory cells' retention. Twelve taps on one chip (a different chip than was tested for Fig. 9.10) were programmed to voltages ranging from 0.25 to 0.75 V and monitored for an hour. The drift rates are shown in Table 9.2 and indicate minimal dependence of retention on temperature.

The other common way to store an analog voltage is with a switched-capacitor (SC) sample and hold circuit. Even using large capacitors, SC sample and hold circuits tend to experience far more rapid voltage droop than the floating-gate memory

Table 9.2 Effect of temperature changes on retention

Tap	Drift (μV/s) at 25°C	Drift at 75°C	Tap	Drift (μV/s) at 25°C	Drift at 75°C
1	0.763	−0.058	7	−1.509	−1.022
2	−4.035	−3.109	8	0.000	−1.737
3	−9.103	−7.095	9	−2.339	−2.715
4	−0.661	−0.876	10	−5.153	−5.299
5	−1.949	−2.000	11	1.577	1.066
6	−2.967	−2.073	12	−5.374	−4.642

cell described here. For example, the circuit in [5] uses a 100 pF capacitor and has a droop rate of 10 V/s.

On the other hand, floating-gate memory cells using thick-oxide devices achieve extremely good retention, with droop rates below 1 mV in 10 years [1]. For an application where the memory cells are part of a continuously adapting system, the retention demonstrated here is sufficient. To avoid degradation of clustering performance due to memory cell leakage, the voltage change due to leakage should be much smaller than that due to intentional updates. The data in Fig. 9.9a shows a minimum update rate of 20 mV/s with $V_{ReF} = 0.8\,\text{V}$. If we use a 2 ms update pulse, and spikes occur at a rate of 10 spikes/s, then a centroid receiving all of its updates in the same direction (i.e., one that had not yet converged), would accumulate updates at a rate of about 400 µV/s, or slightly more than ten times the worst case drift seen in Fig. 9.10.

9.4.3 Classification

Every point presented to the clustering circuit is classified based on its proximity to the four centroids. Classification depends on the distance measurement performed by the Gilbert cell squared difference circuit and loser-take-all decision circuit. Figure 9.11 shows the output current from 12 cells. The current output saturates for input differences larger than about 150 mV. Additionally, a significant variation in the maximum output current can be seen in the figure. As a result, classification decisions involving two centroids that are both far from the received data point will be resolved based on the maximum output currents of the respective multipliers rather than on the actual distance between the centroids and the input.

There is also variation in the input voltage that elicits the minimum current from the difference circuits, due to threshold variation in the input transistors of the Gilbert cell. For classification purposes, the centroids can be considered to be located at the sum of the memory cell values and the difference circuit offsets. Tables 9.3 and 9.4

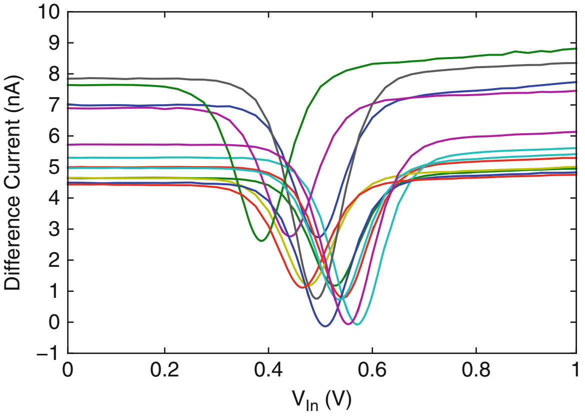


Fig. 9.11 Output current (top) from the squared-difference circuits for 12 cells on a single chip as a function of input voltage, for cells programmed to 0.5 V

Table 9.3 Input offset voltage, maximum current, minimum current for distance circuit

	Class 1	Class 2	Class 3	Class 4
Dim 1	29.6 mV	17.4 mV	45.3 mV	71.2 mV
	7.73 nA	4.94 nA	5.29 nA	5.61 nA
	2.74 nA	1.18 nA	0.79 nA	−0.08 nA
Dim 2	−57.9 mV	−67.1 mV	−3.0 mV	−4.2 mV
	7.45 nA	5.01 nA	8.35 nA	4.83 nA
	2.76 nA	1.19 nA	0.76 nA	−0.13 nA
Dim 3	−71.4 mV	−36.3 mV	21.8 mV	45.2 mV
	8.81 nA	4.75 nA	5.41 nA	6.13 nA
	2.62 nA	1.11 nA	0.73 nA	−0.06 nA

Table 9.4 Comparator offset voltage

	Class 1 (mV)	Class 2 (mV)	Class 3 (mV)	Class 4 (mV)
Dim. 1	−24.5	14.1	−26.3	−2.9
Dim. 2	42.7	36.0	50.3	84.2
Dim. 3	13.5	−16.6	−40.1	0.7

summarize parameter variation across the difference circuits and memory cell comparators, respectively.

Finally, the conversion from a differential current to a single-ended current creates variation in the minimum output current. Because the differential-single-ended conversion occurs after the 1D distance currents are summed for each class, the current minimum is common to each of the three cells in each class. Close inspection of Fig. 9.9a reveals that there are four groups of three curves, with each group having a common minimum.

After the one-dimensional difference currents are summed across dimensions for each class, the loser-take-all determines which class is nearest to the input and outputs a binary vector with one bit for each class. Figure 9.12 shows the result of sweeping the three input voltages together while the stored voltages remain constant. The binary indicator outputs for each of the four classes are shown in the top four traces (labeled W1–W4). The bottom plot shows the difference currents for the four classes. For large differences between the distance currents, only one indicator is high. For inputs where the smallest currents are similar, more than one indicator may be high, effectively classifying an input as belonging to multiple classes. In the case shown here, there is an overlap between W1 and W2 when the difference between the respective currents $I_{Diff,1} - I_{Diff,2}$ is between −1.9 and 2.9 nA.

The distance current and loser-take-all circuits combine to provide a classification label for every input during clustering. Figure 9.13 shows the classification results for a two-dimensional slice of the input space for one set of centroid locations. The centroid locations are represented by the diamonds. The centroids' third dimensions are programmed to 0.5 V plus the input-referred offset of their respective distance circuits to minimize the contribution of the third dimension to the output distance current for a constant input of 0.5 V. Each diamond is accompanied by a short line

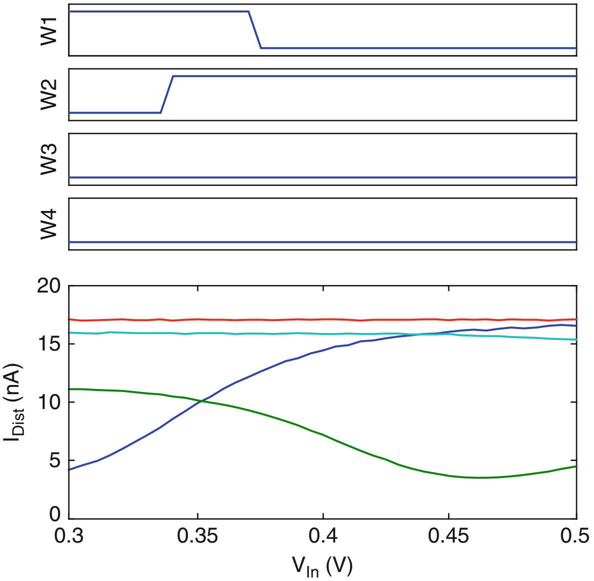


Fig. 9.12 The three input voltages are swept together while the stored voltages remain constant. The binary indicator outputs for each of the four classes are shown in the top four traces. The bottom plot shows the four difference currents. For large differences between the difference currents, i.e., when current is clearly smaller than the others, only one indicator is high. For inputs where the smallest currents are similar, more than one indicator can high

indicating the movement due to leakage of the centroid during the experiment. Centroids 2–4 did not move a noticeable amount, and class 1 moved only slightly (about 20 mV). The areas labeled by each class are outlined. Points labeled with multiple classes result in an overlap of the class boundaries.

The distance currents for each of the four classes were measured at the same time as the classification results and are shown in the four surface plots in Fig. 9.14.

9.4.4 Clustering Convergence

To demonstrate the functionality of the memory cell and classification circuits working together, different data were presented to the clustering circuit under various conditions.

Figure 9.15 shows measured data demonstrating the convergence of the clustering circuit for the trivial case of a single cluster centered at 0.5 V in all three dimensions. In this test, the winner and loser pulse widths were both set to 25 ms, so all of the centroids adapted based on the presented data, irrespective of the classification results. This demonstrates that the memory cells successfully converge to the location of a cluster of input data.

Figure 9.16 shows the results of learning two clusters of data with four centroids. All four centroids were initialized to approximately (0.5 V, 0.5 V, 0.5 V) in their three dimensions. It can be seen that the clustering circuit is able to separate the two classes, with data in the bottom left assigned to classes 1 and 3 and the top right assigned to classes 2 and 4. While a few points in the far upper-right include a blue

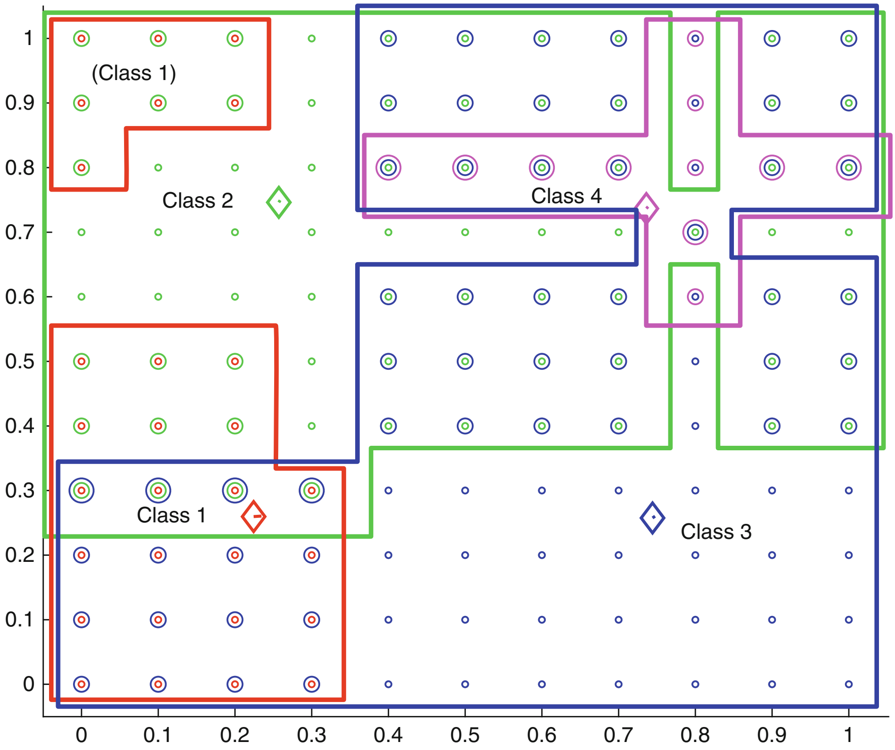


Fig. 9.13 Classification results for a 2D slice of the input space. Classes one through four are represented by the diamonds in the bottom left, top left, bottom right, and top right respectively. Each classified data point contains a circle for each class that output a '1' for that point, so points that were labeled as belonging to multiple classes have multiple concentric circles, colored according to the classes

(class 3) label, none in the bottom-left include a label of class 2 or 4. With a few known samples and the corresponding classification results, it should be possible for the multiple labels to be interpreted correctly with fairly simple post-processing. Figure 9.16b–d shows each dimension separately.

Here the interaction of classification and centroid adaptation can be seen. For example, class 2 (x), initially labels some of the lower cluster (centered around 0.3 in all three dimensions) and begins adapting downwards. However, class 3 adapts more quickly, and soon begins labeling the lower clusters exclusively around iteration 200. At this point, class 2 is also labeling some points in the upper cluster and reverses direction, heading upwards. After about 750 points, class 2 is exclusively labeling most of the upper cluster.

One important performance metric for clustering implementations is the ability to separate closely spaced clusters of data. Figure 9.17 shows the results of three experiments designed to evaluate the resolution of the clustering circuit. In Fig. 9.17a, two clusters separated by 200 mV are successfully resolved. In Fig. 9.17b, the

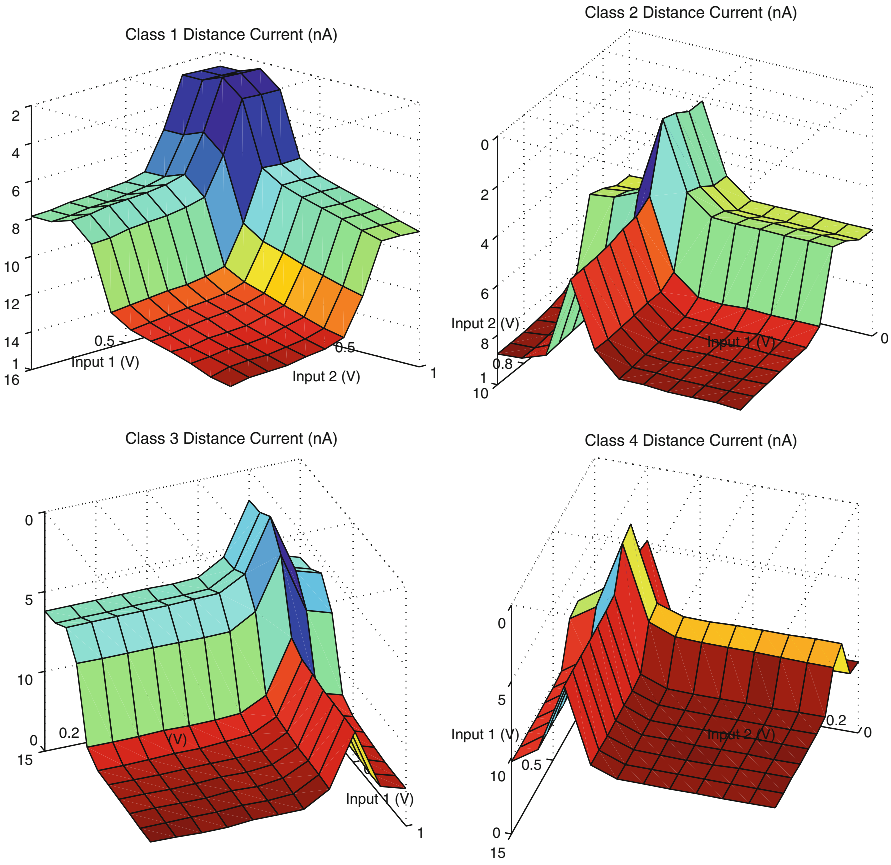


Fig. 9.14 Surfaces showing the distance current as a function of two input variables. The distance currents shown here correspond to the classification results shown in Fig. 9.13. Note that the orientation of the z-axis (distance current) is reversed, so that the highest points on the surface correspond to the points nearest the centroid. Also note that the x and y-axes are rotated to best show the surface, and are not all shown from the same angle

separation is reduced to 100 mV, and the circuit is unable to distinguish the two clusters.

This limitation comes about through the interaction of the saturating nature of the distance measurement performed by the Gilbert cells, the finite gain in the LTA circuit, which allows multiple classes to equally claim a single data point, and the adaptation asymmetry. Because of the multiple-labeling behavior of the LTA, the two centroids near the data are both claiming every point. As a result, they are both attempting to learn the center of the *entire* data set, rather than the center of just one cluster. Because of the update asymmetry, the two centroids actually converge to a point on the periphery of the data set rather than the actual center, as explained in the previous chapter. The centroid's location on the periphery of the data set causes most

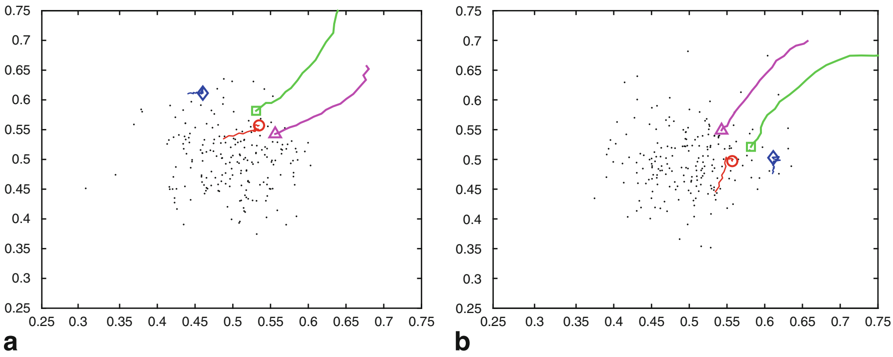


Fig. 9.15 Four centroids in 3D learning one cluster. Classes 1–4 are marked with the diamond, square, circle, and triangle, respectively. Dimensions one and two are shown in (**a**), while dimensions two and three are shown in (**b**)

of the inputs to land in the saturated, low-gain, region of the Gilbert cells, preventing either centroid from exclusively labeling many points, and thus preventing the natural positive feedback of the clustering algorithm from taking effect.

In Fig. 9.17c, the reference voltage is increased from 0.8 to 0.9 V, improving the symmetry for high-valued inputs. This improved symmetry allows the two centroids to get closer to the actual center of the data, bringing the inputs into the effective range of the distance circuits, and allowing the classes to begin exclusively claiming inputs. The two clusters are successfully resolved. It should be noted that the increase in V_{Ref} would actually cause the situation to become worse for lower-valued inputs.

Figure 9.18 shows classification results for spikes from a synthesized neural recording. In this experiment, adaptation was disabled, and the centers were pre-programmed to the desired locations in order distinguish two different spike shapes. Because the classification output is a 4-bit vector, there are 16 possible labels, of which 5 are present in the results from this experiment. The bottom-left cluster was labeled exclusively by class 3, while the top-right cluster was labeled primarily by class 2, but had some points with multiple labels. Despite the multiple labelings, there is clearly sufficient information in the classification results to distinguish the two clusters. The corresponding spike waveforms, shaded according to the label combinations are shown in Fig. 9.18b.

9.5 Discussion

This chapter described a floating-gate memory cell and its application in a clustering circuit. The circuit was able to successfully cluster well-separated data. Circuit non-idealities limit the resolving power of the circuit to the extent that the realized implementation is not suitable for autonomous general-purpose spike sorting. However,

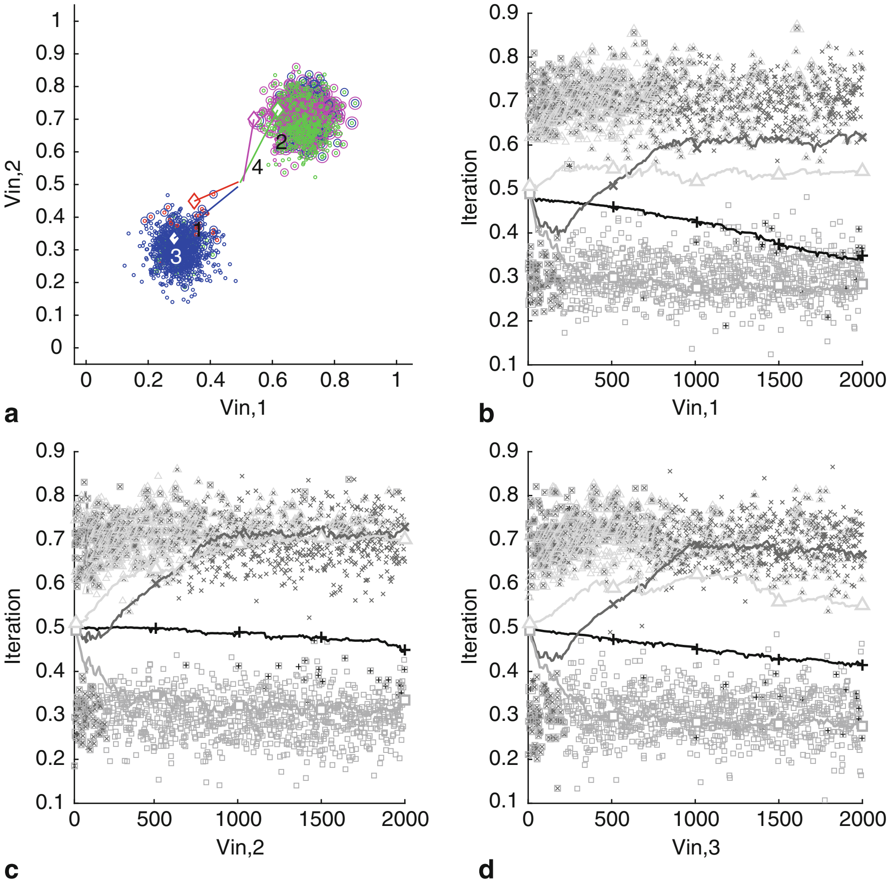


Fig. 9.16 Clustering of a data set with two clusters in three dimensions. **a** The first two dimensions. The diamonds show the four centers, with tracks indicating movement from the initial positions. Each point is marked with a circle for each class in which it is included. **b–d** Dimensions 1–3 respectively, with inputs and centroid locations plotted against time. Class 1 = black +, Class 2 = dark gray x, Class 3 = medium gray square, Class 4 = light gray triangle. Points labeled by multiple classes have corresponding multiple markers, so a point labeled by classes 2 and 4 has a dark gray x inside a light gray triangle

the performance of the classification component was demonstrated to successfully resolve spike shapes in a non-adaptive setting.

Previous implementations of learning algorithms with floating-gate circuits have used thicker oxides to enable long-term retention. The work described here is the first work use of such thin gate oxides (2.2 nm physical thickness) in floating-gate-based learning circuits. The use of thin oxides allows the circuit to operate with a supply voltage of about 1.5 V, enabling very low power dissipation, and operation from a small battery.

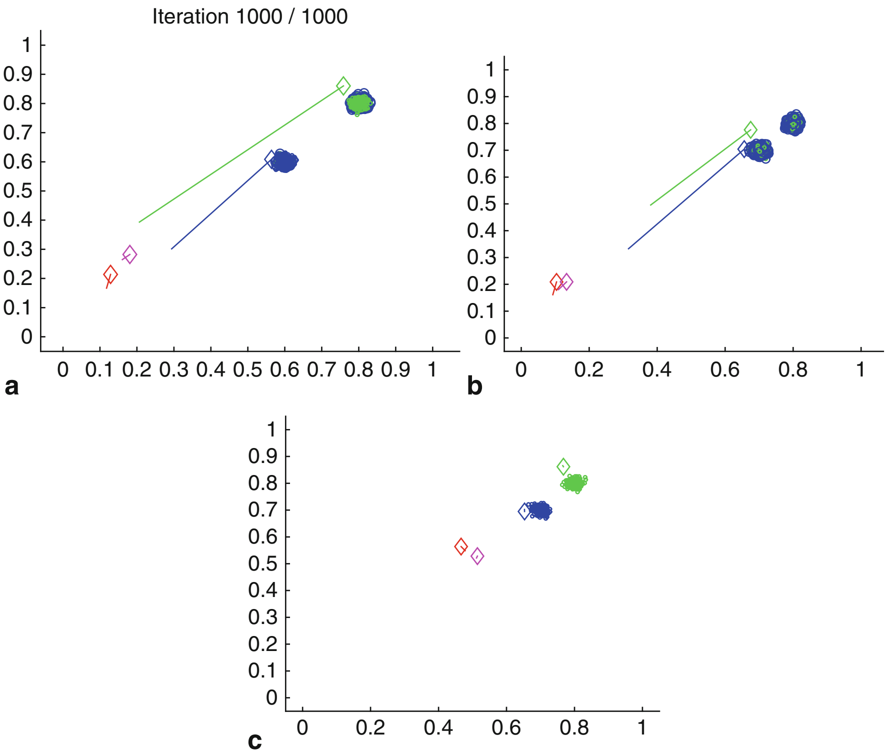


Fig. 9.17 Evaluation of the ability of the clustering circuit to resolve closely spaced clusters. **a** Two clusters with means separated by 200 mV are successfully resolved. **b** With the separation reduced to 100 mV, the circuit fails to separate the two clusters. **c** With the V_{Ref} increased from 0.8 to 0.9 V, the improved symmetry of the adaptation allows the two clusters to be distinguished

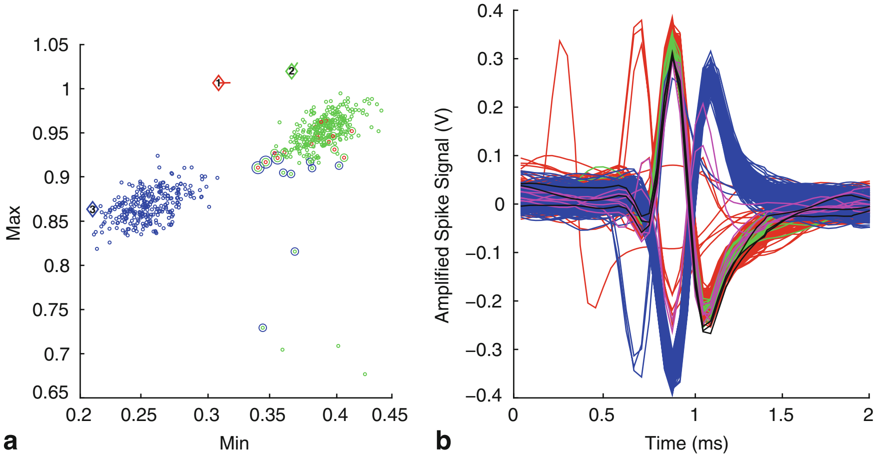


Fig. 9.18 Classification of spikes with pre-programmed centroid locations. **a** Classification results in the first two dimensions of the feature space (minimum and maximum voltage of the spikes). **b** Spike waveforms colored by classification

Additionally, the results shown here point to potential circuit modifications that should yield significantly improved clustering performance. Because the limited gain of the LTA decision circuit and the resulting multiple-labels undermined clustering performance, the LTA circuit should be designed to guarantee that only one output can be high. It is likely that positive feedback could be triggered when the inputs are sampled to increase the effective gain. The dependence of adaptation rates on the stored value prevented update symmetry from being simultaneously achieved across the input range. This could easily be addressed in future versions by leaving feedback intact during adaptation, which would keep the floating gate at V_{Ref} during tunneling. Circuit techniques to automatically balance the update rates should also be investigated.

Beyond performance improvements to the circuit, analog machine learning algorithms would also benefit from more flexible architectures. The clustering circuit described here can only operate with three input dimensions and four classes. Relatively simple changes could allow the number of classes and dimensions to be adjusted as needed. Flexible architectures, similar to those described in [2] would also make analog computation more attractive for deployment in real systems.

References

[1] Ahuja B, Vu H, Laber C, Owen W (2005) A very high precision 500-na CMOS floating-gate analog voltage reference. IEEE J Solid-State Circuits 40(12):2364–2372

[2] Bridges S, Figueroa M, Hsu D, Diorio C (2005) A reconfigurable VLSI learning array. In: Proceedings of 31st European conference on solid-state circuits (ESSCIRC 2005). Grenoble, France, pp 117–120

[3] Delbruck T (1991) Bump circuits for computing similarity and dissimilarity of analog voltages. In: Proceedings of international joint conference on neural networks, July 8–12, Seattle, Washington, pp 475–479

[4] Figueroa M, Bridges S, Diorio C (2005) On-chip compensation of devicemismatch effects in analog VLSI neural networks. In: Saul L, Weiss Y, Bottou L (eds) Advances in neural information processing systems, vol 17. MIT Press, Cambridge, MA, pp 441–448

[5] Gasparik F (1980) A precision autozeroing sample and hold integrated circuit. IEEE J Solid-State Circuits 15(6):945–949

[6] Hsu D, Figueroa M, Diorio C (2001) A silicon primitive for competitive learning. Adv Neural Inf Process Syst 13:713–719

[7] Hsu D, Bridges S, Figueroa M, Diorio C (2003) Adaptive quantization and density estimation in silicon. Adv Neural Inf Process Syst 15:1107–1114

[8] Lenzlinger M, Snow EH (1969) Fowler-nordheim tunneling into thermally grown SiO_2. J Appl Phys 40(1):278–283

[9] Patel G, DeWeerth S (1995) Compact current-mode loser-take-all circuit. Electron Lett 31(24):2091–2092

Chapter 10
NeuralWISP: A Wirelessly Powered Spike Density Recording System

Having discussed the critical components of implantable neural interfaces, we may now move on to system integration issues. After reviewing some previous system-level efforts, this chapter and the next will describe two example systems, both including signal acquisition circuitry and a wireless communication link.

10.1 Previous Neural Recording Systems

Most successfully deployed neural interfaces to date have used discrete commercial ICs on a custom printed-circuit board. The NeuroChip [10, 11] developed by Mavoori et al. has been successfully used to enable novel neuroscience experiments in primates [9]. It includes amplification, digitization, storage in on-board flash memory, local processing using a microcontroller, and an infra-red interface. Local spike discrimination is performed using user-defined time-amplitude window discriminators.

Santhanam et al. developed HermesB, a recording system featuring amplification, digitization, processing with an ARM microprocessor, motion measurement by an on-board accelerometer, and local storage in flash memory. Spike sorting is performed locally using the Sahani algorithm [13].

The advantages of this type of architecture include relatively fast prototyping (compared to custom IC design), and flexibility due to the ability of the microcontroller to run arbitrary software. The disadvantage is the large size and high power consumption. The neurochip is $1.2 \times 5.4\,\text{cm}^2$ and consumes 20–60 mW, depending on configuration. Including batteries, the total implant size is $5.5 \times 5 \times 3\,\text{cm}^3$ and weighs 56 g. HermesB is $6 \times 7 \times 4.5\,\text{cm}^3$, and weighs 220 g including the aluminum enclosure, batteries, and hardware associated with the electrode array. It consumes 70–320 mW from its batteries depending on operating mode.

The size of such devices precludes implantation under the skull. The development of highly integrated neural signal processors should enable a reduction in the physical size of neural interfaces both by reducing the number of components needed and by reducing the battery requirements.

DOI 10.1007/978-1-4419-6727-5_10, © Springer Science+Business Media, LLC 2011

In an effort to reduce the physical size and power requirements for neural interfaces, researchers have developed custom integrated circuits. In [5], Harrison et al. describe an Integrated Neural Interface (INI) for a 100-electrode recording system, which includes amplifiers, threshold-based spike detection, a single user-selected channel of full-waveform digitization, and a wireless transceiver. It utilizes $4.7 \times 5.9\,\text{mm}^2$ of chip area and consumes 13 mW from 3.5 V, supplied by an inductive link. Because the INI is powered from an external source, and does not include storage or local processing, it is not fair to directly compare area and power to more complete systems such as the NeuroChip or HermesB. However, it does provide a hint of the potential advantages of aggressive integration. In [2], a subsequent generation of the INI chip is integrated with an electrode array, a battery and and antenna to perform recording and wireless transmission of spiking activity. The integrated system consumes 46.8 mW from a 4 V supply.

In [1] Chae et al. reported a 128-channel neural recording IC including amplification, integrated spike detection and feature extraction implemented with on-chip DSP circuitry, and a wireless transmitter. Their recording IC consumes 6 mW from a 3.3 V supply. The detection and feature extraction blocks consumed approximately 1 μW/channel. The extracted features are the maximum and minimum of each detected spike. Features are transmitted off-chip through the wireless interface for spike sorting or other analysis.

Most previous implantable neural recorders have used a simple threshold for spike detection and have included no spike sorting facility at all. One exception is [12], which transmitted spike amplitude off-chip for use in off-line sorting. However, Olsson and Wise's system digitized the entire waveform, thus incurring the cost of constant analog-digital conversion. We are not aware of any implantable system designed to date that includes local spike sorting, or of any analog spike-sorting system.

Because transcutaneous wiring poses a significant infection risk, it is desirable that a neural interface communicate and receive power wirelessly. Previous systems [6, 12] have achieved wireless operation by using a near-field inductive link to transmit power and data. However, these systems require that the external coil be located within a few centimeters of the internal coil. A wireless neural interface with a range of 1 m or more will enable the removal of the interrogator from the head and would allow wireless interfaces to be placed on small animals incapable of carrying the interrogator hardware, such as mice.

This chapter describes a wireless neural interface [8] which harvests power from the radio-frequency (RF) energy provided by a standard commercial UHF RFID reader. Figure 10.1 illustrates the use of the NeuralWISP. The system operates at a distance of up to 1 m from the reader. It records the number of spikes that occurs in a programmable window (typically 1–10 s) and subsequently transmits the spike count to the reader as part of the tag identification number that the reader is designed to acquire. This allows the neuroscientist a wireless, battery-free method of recording spike density ($\frac{spikes}{second}$) as various tasks are performed or stimuli are presented.

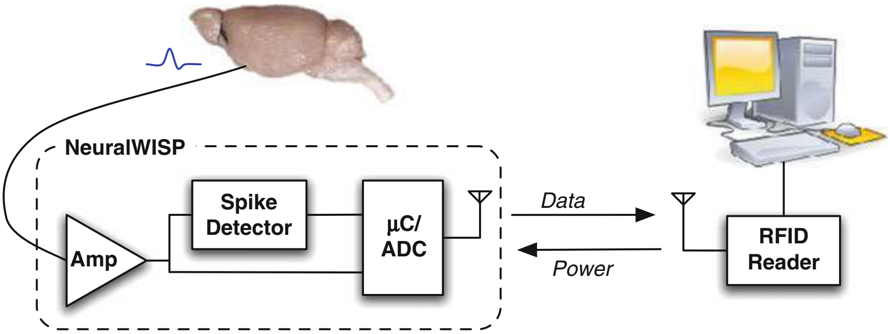


Fig. 10.1 Usage model of the NeuralWISP

10.2 System Design

The NeuralWISP is an extension of the Wireless Identification and Sensing Platform (WISP) [14, 16]. The WISP is a fully-passive UHF RFID tag that uses an low power, 16-bit, general-purpose microcontroller (µC) for sensing, computation and RFID communication. The use of a programmable µC allows WISP to be easily configured for different applications including measurement of temperature, light level, strain, and acceleration [16]. In monitoring applications, analog sensor outputs change slowly and thus permit periodic, low-frequency (1–50 Hz) measurement. However, a much faster sampling rate (at least 8 kHz) is necessary to detect neural spikes. Achieving this sampling rate under the constraints of the limited power budget of an RFID tag is not possible with general purpose microcontrollers available today.

In order to minimize the average current consumption, a continuous-time analog spike detector was designed to generate a µC interrupt when a spike occurs. This allows the µC to remain in a low-power sleep mode during periods of inactivity and only wake up to process spikes or communicate with the RFID reader. The µC counts spikes during a programmable window and is reset after the spike count is transmitted to the reader.

The architecture of the NeuralWISP is shown in Fig. 10.2. Like a typical RFID tag, power is received at the antenna, voltage-multiplied, rectified, and regulated to provide a stable system power supply. The amount of power received is a strong function of wireless range as modeled by Friis' Transmission Equation. To illustrate the extremely limited power budget, a graph of available power (after rectifier losses and quiescent current draw) is shown in Fig. 10.3. The design and performance of this energy harvesting circuitry is described in detail in [14, 16]. The neural input signal is amplified and applied to an analog spike detector in addition to an analog-digital converter (ADC) integrated in the µC. The µC performs the control and timing tasks, and implements the RFID communication protocol.

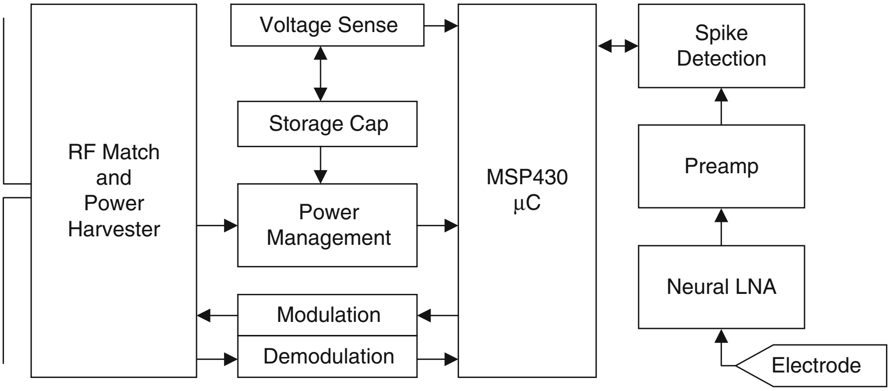


Fig. 10.2 Block diagram of NeuralWISP

10.2.1 Analog Signal Path

The extremely low signal levels recorded from neural probes place severe constraints on the analog front-end. Input-referred noise levels must be $<10\,\mu V_{RMS}$ while providing good linearity and high gain. These requirements frequently result in the low noise neural amplifier consuming a majority of the system power. In the NeuralWISP, the power dissipation limits the wireless range, so power must be minimized. A custom low-noise amplifier (LNA) was designed in a 0.5 µm SOI BiCMOS process to meet these requirements. The amplifier is designed to provide a gain of 40 dB. A schematic is shown in Fig. 10.4.

The amplifier is built using a two-stage op-amp with capacitive feedback. A closed-loop configuration was chosen for this system because open-loop amplifiers, while demonstrating superior noise efficiency factors (NEF), typically suffer from inferior power-supply rejection [7]. MOS-bipolar pseudo-resistors [4] (PR) were

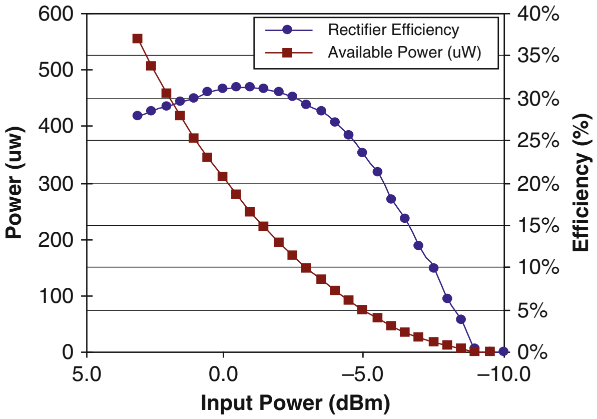


Fig. 10.3 Rectifier output power and efficiency versus input power. Note that 0 dBm corresponds to approximately 1 m for a typical UHF RFID system

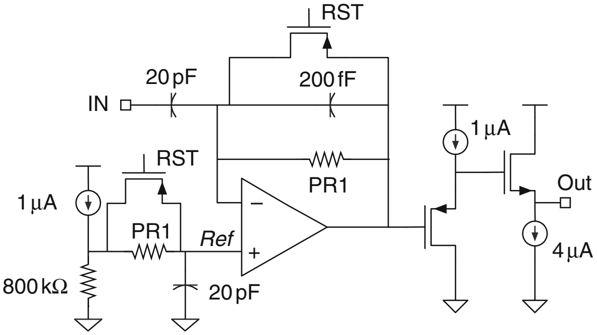


Fig. 10.4 Schematic of custom 8 µA low noise neural amplifier fabricated in a 0.5 µm SOI CMOS process

used to set a sub-Hz low frequency pole for DC rejection. For small signals, the PRs have an incremental resistance of about 10^{12} Ω, resulting in a time constant of several seconds. In order to avoid long settling times on power up, a power-on-reset circuit is included on chip which temporarily shorts out the pseudo-resistors. Reset can be driven from an external pin, so it could also be used to speed recovery from stimulation artifacts. The high-pass corner frequency set by the pseudo-resistors is much lower than is necessary for the extra-cellular recording task being demonstrated here. However, transconductor implementations of high-valued resistors consume additional power and contribute noise.

A source-follower output stage was chosen for its flexibility with respect to load conditions. A resistive load to ground will increase the current in the NMOS source follower transistor, allowing the amplifier to automatically adapt to resistive loads without consuming extra static bias current under lightly loaded conditions or using a complicated class AB output stage. The chip is completely self-contained, and includes a supply-independent bias current generator allowing consistent operation over a range of 1–5 V.

An additional gain of 20 dB is provided by a second amplifier built from two OPA349 op-amps, shown in Fig. 10.5. The first opamp is used to establish a 0.6 V reference for AC coupling the amplifier stages, and the second opamp is used in a non-inverting gain configuration. The gain of the first stage allows relatively noisy micro-power op-amps to be used for the second gain stage. Consequently, the second stage consumes only 1.9 µA from a 1.8 V supply, including the reference.

Figure 10.6 shows the noise spectra at the output of both the LNA and the post-amp. Even with the use of the micro-power commercial op-amps, it can be seen that the gain of the LNA suppresses the noise contribution of the post-amp. Additionally, some very low-frequency noise is filtered by the AC coupling between the LNA and post-amp.

The output of the second amplifier is connected to the ADC input of the MSP430 microcontroller to allow for direct digitization of the neural signal. Additionally, the amplified signal is applied to an analog spike detector. The signal is low-pass filtered with a time constant of 0.1 s to generate the detection threshold. The signal is also

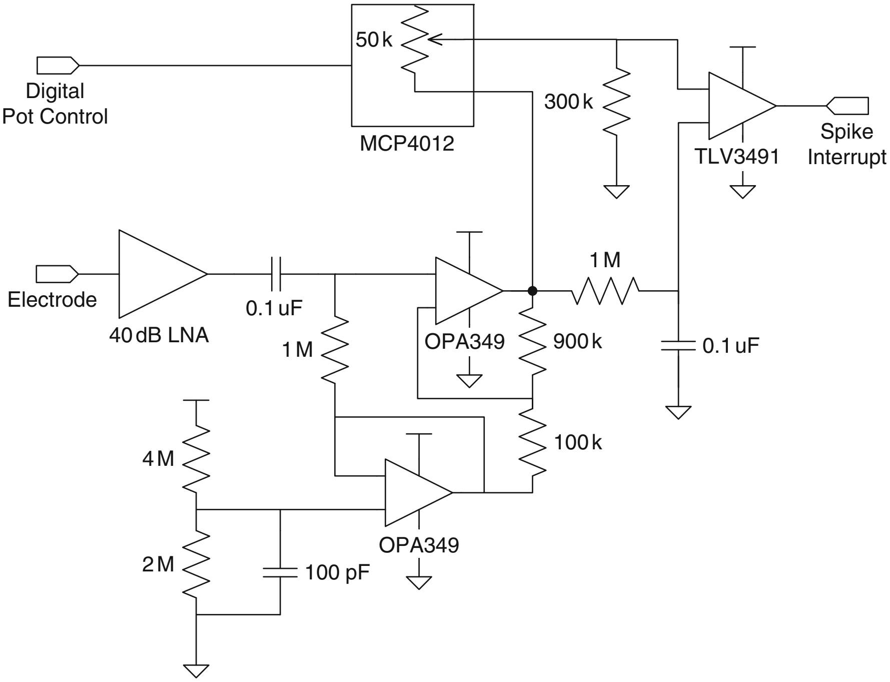


Fig. 10.5 Analog front end circuitry, including custom LNA, 20 dB post-amp, and spike detector with programmable threshold

attenuated and shifted towards 0 V by up to 15% via a variable-ratio resistive divider. A digitally-controlled resistor, variable from 0 Ω–50 kΩ, determines the attenuation of the divider and thus the sensitivity of the spike detector. The spike detector's programmable threshold is set by the µC, allowing adjustment for dynamic neural signals and noise levels.

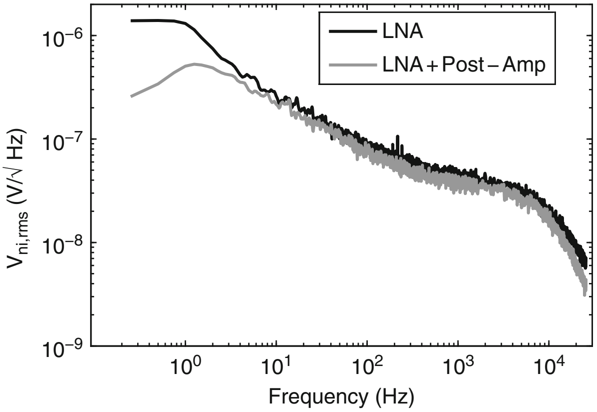


Fig. 10.6 Noise spectra at the output of both the LNA and the post-amp

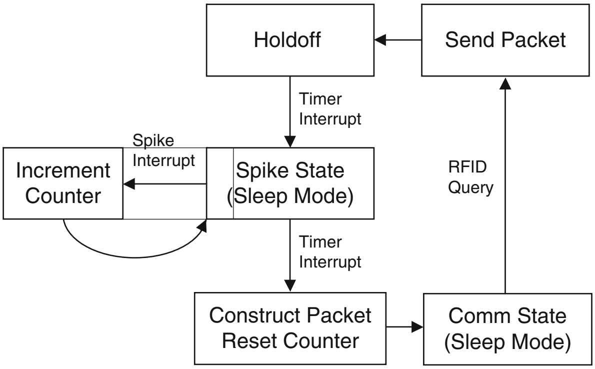


Fig. 10.7 Software state diagram. The µC is in the low-power Spike State for the majority of the time, awakening only to increment the spike counter after a detection or to communicate with the reader

10.2.2 Digital Control

An MSP430F2274 microcontroller (µC) is used to implement control, timing, and communication tasks. Figure 10.7 shows the software architecture. On boot-up, the µC configures the adjustable resistor in the spike detector. During the primary mode of operation, the µC will count spikes during a user-specified time interval (typically 1–10 s) and transmit the number of spikes detected at the end of the interval. During the counting interval, the µC is in a low-power sleep state for the majority of the time. The spike detector triggers an interrupt, which causes the µC to wake up, increment the spike count, and return to sleep. A timer drives another interrupt, which signals the end of the counting interval, causing the µC to exit the spike-counting mode and await a communication session with the reader. After communicating with the reader, the µC pauses for 1 s to allow the analog circuits to recover from RF interference that occurred during the read, then returns to the spike counting phase and repeats the cycle.

10.3 Test Results

The fabricated board is shown in Fig. 10.8. The populated board alone weighs 1.0 g, and a 900 MHz wire dipole antenna (not shown) weighs approximately 0.6 g. During spike counting, the system draws an average of about 20 µA of current from its unregulated supply, of which 8 µA is consumed by the neural LNA. A commercial RFID reader with +30 dBm transmitted power was used to wirelessly supply power and communicate with the NeuralWISP.

The input-referred noise of the low-noise amplifier is 4.4 μV_{RMS}, measured from 0.25 Hz to 25 kHz. Operating from a supply between 1 and 5 V, the LNA provides a measured gain of 39 dB with a bandwidth spanning 0.5 Hz to 5.9 kHz. Current consumption at 1.8 V is 8 µA, including the bias generator and output buffer. Figure 10.9 shows the frequency response of the first stage and the combined response of both gain stages. The LNA combined with the second amplifier provides a mid-band gain of 56 dB with a bandwidth from 2 Hz to 4.9 kHz.

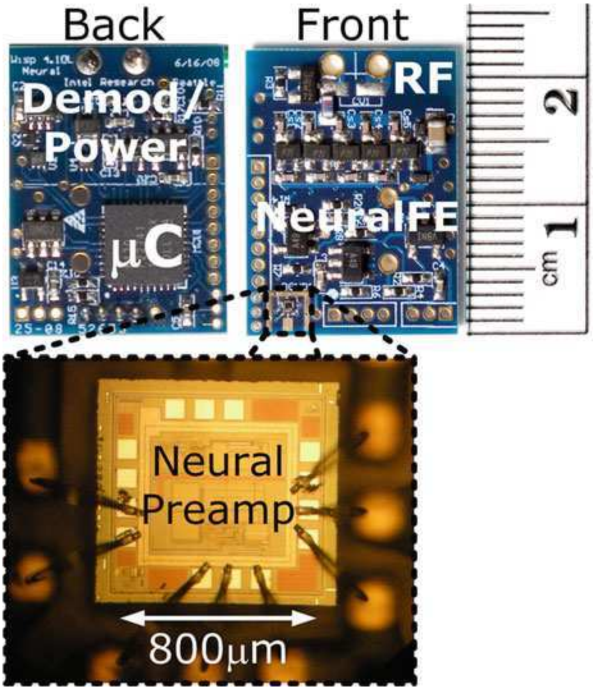


Fig. 10.8 System photograph. Inset shows chip-on-board mounting of the custom low-noise amplifier IC

To characterize the spike detector, we applied a synthesized neural recording [15] to the NeuralWISP input. This technique allowed us to vary the SNR and spike rate in the recording and provided a reference against which to compare our measured spike detection results in order to characterize the detector accuracy. Figure 10.10 shows the operation of the detector on a single spike, with an 800 µV$_{P-P}$ input signal. Software debouncing in the interrupt handler prevents any glitches in the spike detection signal from causing errors in the spike count.

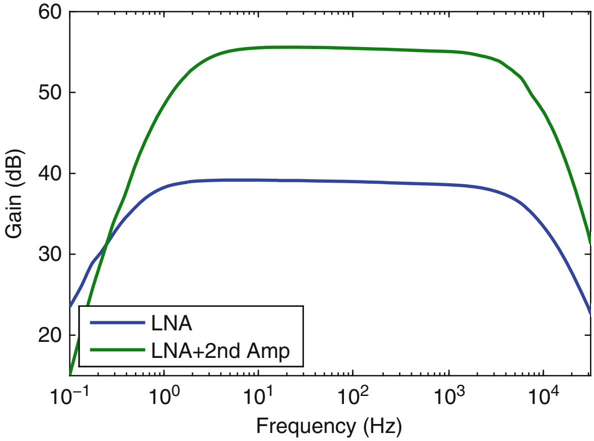


Fig. 10.9 Gain versus frequency for both low-noise amplifier (LNA, bottom), and the combined gain of the LNA and 2nd amplifier

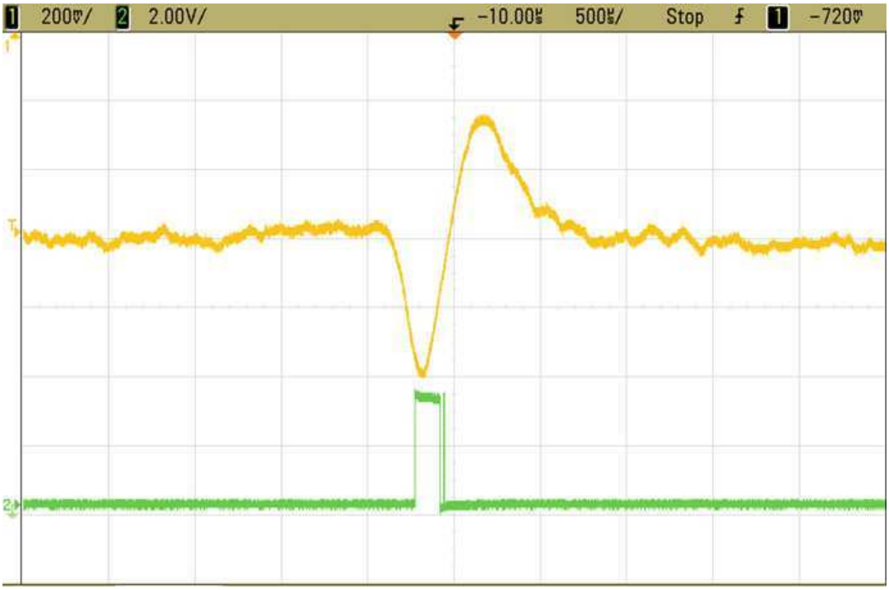


Fig. 10.10 Operation of the spike detector. The input signal (top) has been amplified by ×1000 for oscilloscope viewing. The amplitude at the input to the NeuralWISP is approximately 800 µVpp

Figure 10.11 shows the spike detector accuracy. Spikes were detected using the hardware analog spike detector (circle tick) and also using a PC-based threshold-crossing detector (square tick) for comparison. Both detectors were run on synthetic recordings with an amplitude of approximately 400 µV_{P-P} and SNR of 10 dB (left)

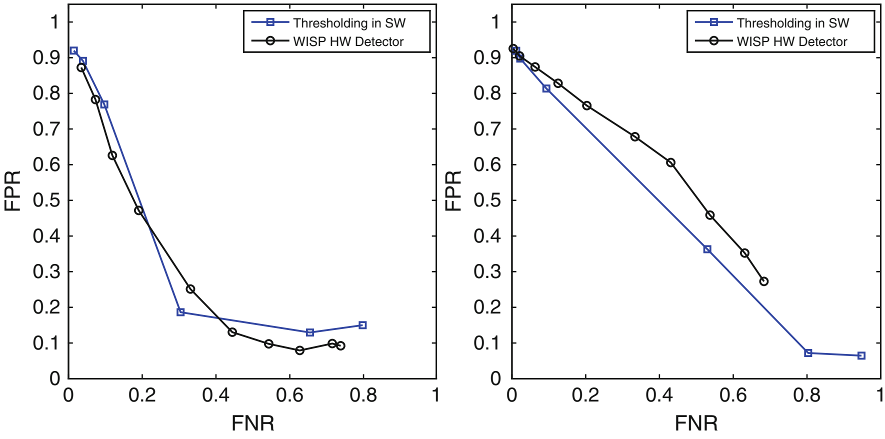


Fig. 10.11 Accuracy of the spike detector compared to a software spike detector for SNR = 10 dB (left) and SNR = 6 dB (right). The x-axis is the false negative rate (FNR = Number of missed spikes / Number of total true spikes). The y-axis is the false positive rate (FPR = Number of false detections / Number of total detections)

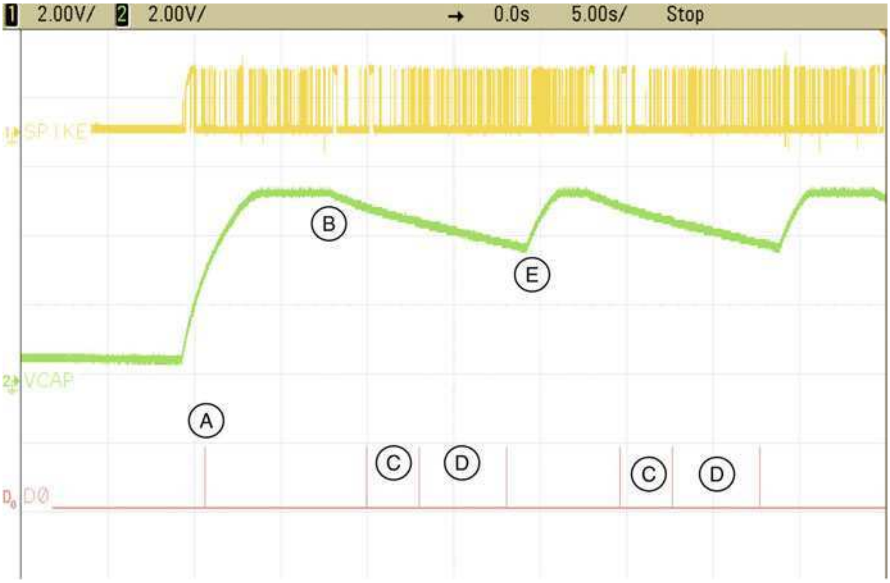


Fig. 10.12 Two read cycles of wireless operation, showing the spike detector output (top), the unregulated stored voltage (middle), and a microcontroller output (bottom) pulsed to show operation. The data was taken at a distance of approximately 1 m from the reader

and 6 dB (right). The results were compared with the known spike times provided by the signal synthesis software. The analog detector demonstrates comparable discriminative abilities to the software detector, indicating that noise contributions from the analog front end do not limit spike detection performance.

Figure 10.12 demonstrates the operation of the NeuralWISP. The middle trace is the unregulated voltage stored on a 100 µF capacitor, which begins at 0 V, since the WISP starts out with no stored energy. As the reader begins to interrogate the NeuralWISP, it operates with the following sequence:

- Initially, the reader is configured to transmit power in continuous-wave (CW) mode, which charges the storage capacitor to 5.5 V where it is clamped by a zener diode.
- (A) As the stored voltage rises, the µC boots up.
- (B) Continuous-wave transmission stops and the RFID reader reads data from the WISP. The first read following bootup will contain empty data.
- (C) Following the read, the µC enters a 3 s waiting state in order to allow the analog circuits to recover from RF interference which occurred during the read.
- (D) After 3 s, the WISP begins counting spikes for 5 s.
- (E) After the spike-counting phase, the reader again transmits CW power to recharge the storage capacitor,
- Another read is executed, which retrieves data from the previous spike-counting phase (D). The cycle is repeated indefinitely.

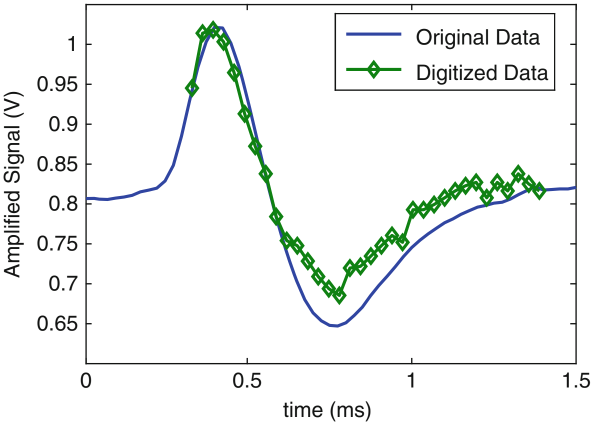


Fig. 10.13 A single spike digitized by the on-board ADC. The μC began sampling and converting in response to an interrupt from the spike detector

The NeuralWISP could also be configured to sample spike waveforms after a spike is detected, and transmit the digitized data. An appropriate duty cycle would need to be chosen in order to meet the constraints imposed by the data rate allowed by the tag/reader interface. Figure 10.13 shows a spike captured and digitized by the NeuralWISP. The digitized spike waveform is superimposed on the original spike waveform. This experiment demonstrates that accurate reconstruction of the spike can be accomplished by waking the μC and ADC from low-power sleep after spike detection, dramatically reducing average system power.

10.4 Experimental Results

To validate the NeuralWISP's ability to detect spikes in vivo, measurements of wing muscle activity from a *Manduca Sexta* moth were taken. While the prototype NeuralWISP is too heavy to be carried by a moth, integration of NeuralWISP onto an IC could allow in-flight measurements to be performed. Because the recording device is wirelessly powered, no batteries or wires are required. Because the battery consumes a large fraction of the weight budget of flying-insect-mounted electronics [3], a wirelessly-powered interface would permit significant weight reduction compared to traditional sensing schemes. The setup is shown in Fig. 10.14, and a wirelessly-powered recording captured by an oscilloscope is shown in Fig. 10.15.

NeuralWISP relies on extremely low-power custom analog front end circuitry to allow operation from a wireless power source. In order to test the compatibility of the analog front end with extra-cellular neural recording, we performed in vivo measurements on a macaque monkey (*macaca nemestrina*). Figure 10.16 shows spikes recorded with the NeuralWISP LNA and post-amp. Standard rack-mounted acquisition equipment was used to digitize the signal and perform spike detection. The signal was filtered with a 4th-order butterworth bandpass filter with bandwidth of 750 Hz–7.5 kHz for spike detection, but the unfiltered signal was stored for offline

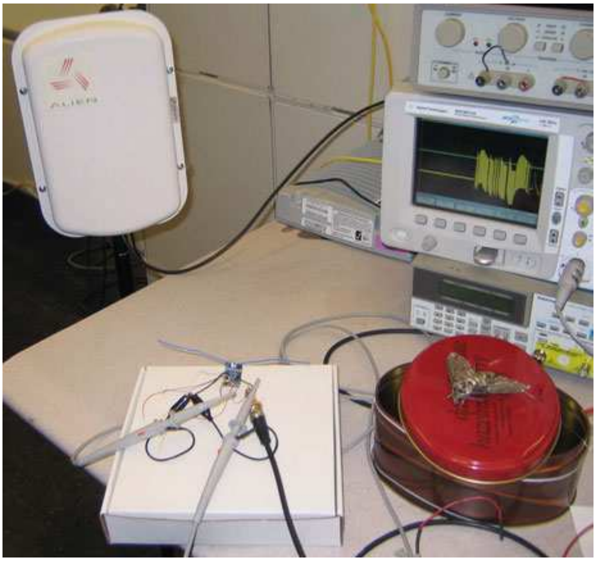


Fig. 10.14 In vivo experiment setup showing *Manduca Sexta* moth with tungsten wire electrodes in wing muscle tissue. The electrodes are connected to the NeuralWISP via a resistive attenuator. Spike density measurements are wirelessly recorded and communicated to the RFID reader

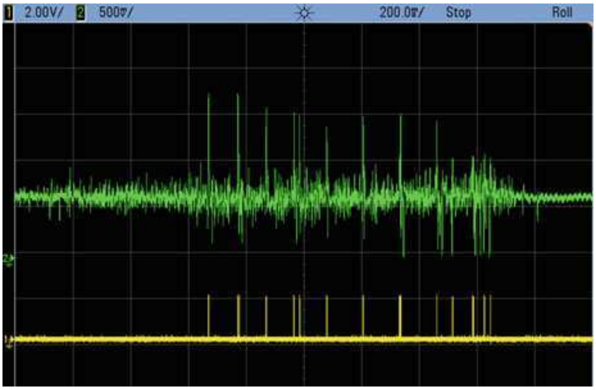


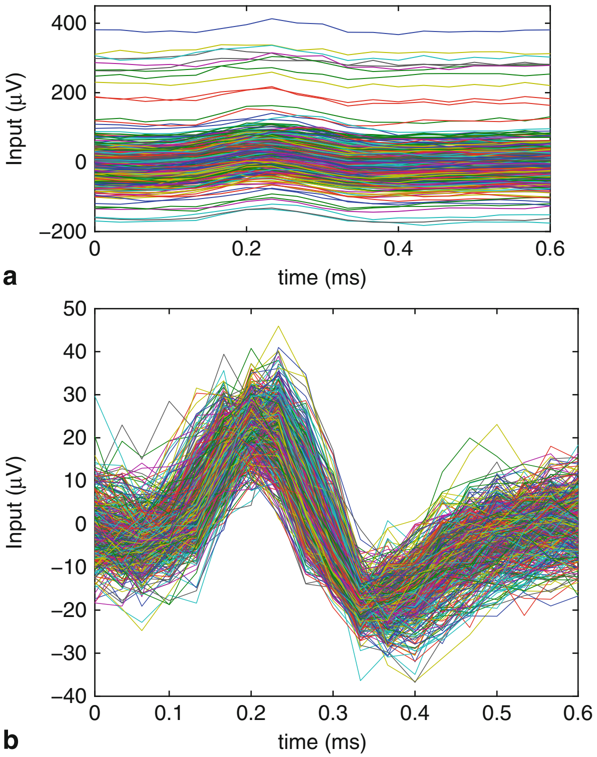
Fig. 10.15 Wirelessly-powered data from wing muscle tissue captured by an oscilloscope. The top trace shows the post-amplifier output; the bottom trace shows NeuralWISP spike detections

processing. Figure 10.16a shows the spikes taken from the raw signals based on timestamps from the acquisition system's detector. Because of low-frequency noise and local field potentials, the spikes are spread widely across the y-axis. In Fig. 10.16b, the same spikes are displayed after filtering with a 750-Hz 2nd-order butterworth high-pass filter.

10.5 Conclusions

Using harvested RF power, the NeuralWISP transmits spike counts to a commercial RFID reader at user-programmable intervals over a range of up to 1 m. In addition to testing with simulation data, in vivo measurements with *Manduca Sexta* moth and macaque monkey validated the feasibility of this system in real-world conditions.

Fig. 10.16 Spikes recorded through the NeuralWISP's amplifiers



By operating from a wireless power source, the NeuralWISP allows indefinite operation without the need to change batteries, a critical need for implanted neural interfaces. The platform is also flexible and can be programmed to operate in different modes, such as spike time-stamp recording, or continuous recording on a duty-cycled basis. Future work reducing the size and weight of NeuralWISP will help lead to the practical deployment of wireless, battery-free neural recording systems.

References

[1] Chae M, Liu W, Yang Z, Chen T, Kim J, Sivaprakasam M, Yuce M (2008) A 128-channel 6 mW wireless neural recording IC with on-the-fly spike worting and UWB transmitter. In: IEEE international conference on solid-state circuits, Digest of Technical Papers, pp 146–147

[2] Chestek C, Gilha V, Nuyujukian P, Ryu S, Shenoy KV, Kier R, Solzbacher F, Harrison R (2008) HermesC: RF wireless low-power neural recording system for freely behaving primates. In: Proceeding of the IEEE international symposium on circuits and systems, pp 1752–1755

[3] Daly D, Mercier P, Bhardwaj M, Stone A, Voldman J, Levine R, Hildebrand J, Chandrakasan A (2009) A pulsed UWB receiver SoC for insect motion control. In: IEEE international in solid-state circuits conference (ISSCC), Digest of Technical Papers, pp 200–201

[4] Harrison R, Charles C (2003) A low-power low-noise CMOS amplifier for neural recording applications. IEEE J Solid-State Circuits 38(6):958–965

[5] Harrison R, Watkins P, Kier R, Lovejoy R, Black D, Greger B, Solzbacher F (2007) A low-power integrated circuit for a wireless 100-electrode neural recording system. IEEE J Solid-State Circuits 42(1):123–133

[6] Harrison R, Watkins P, Kier R, Lovejoy R, Black D, Greger B, Solzbacher F (2007) A low-power integrated circuit for a wireless 100-electrode neural recording system. IEEE J Solid-State Circuits 42(1):123–133

[7] Holleman J, Otis B (2007) A sub-microwatt low-noise amplifier for neural recording. In: 29th annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBS 2007), Lyon, pp 3930–3933

[8] Holleman J, Yeager D, Prasad R, Smith J, Otis B (2008) NeuralWISP: An energy-harvesting wireless neural interface with 1-m range. In: IEEE biomedical circuits and systems conference, Baltimore, MD, pp 37–40

[9] Jackson A, Mavoori J, Fetz E (2006) Long-term motor cortex plasticity induced by an electronic neural implant. Nature 444:56–60

[10] Jackson A, Moritz C, Mavoori J, Lucas T, Fetz E (2006) The neurochip BCI: towards a neural prosthesis for upper limb function. IEEE Trans Neural Syst Rehabil Eng 14(2):187–190

[11] Mavoori J, Jackson A, Diorio C, Fetz E (2005) An autonomous implantable computer for neural recording and stimulation in unrestrained primates. J Neurosci Methods 148(1):71–77

[12] Olsson R, Wise K (2005) A three-dimensional neural recording microsystem with implantable data compression circuitry. IEEE J Solid-State Circuits 40(12):2796–2804

[13] Sahani M (1999) Latent variable models for neural data analysis. PhD thesis, California Institute of Technology

[14] Sample A, Yeager D, Powledge P, Smith J (2008) Design of an RFID-based battery-free programmable sensing platform. IEEE Trans Instrum Meas 57:2608–2615

[15] Vogelstein R, Murari K, Thakur P, Diehl C, Chakrabartty S, Cauwenberghs G (2004) Spike sorting with support vector machines. Conf Proc IEEE Eng Med Biol Soc 1:546–549

[16] Yeager D, Sample A, Smith J (2008) WISP: a passively powered UHF RFID tag with sensing and computation. In: Syed MI, Ahson A (eds) RFID handbook: applications, technology, security, and privacy. CRC Press, Florida

Chapter 11
A 500 µW Wireles Neural Streaming System

This chapter describes a fully integrated neural interface which wirelessly streams a digitized neural waveform over 15 m [2]. In contrast to the NeuralWISP described in the previous chapter, this system is designed to operate from a small battery. The battery allows the system to operate at a greater range from the receiver and to transmit more data. Because of the low power consumption in the analog front end and the wireless transmitter, the system can operate from a very small battery, resulting in an extremely lightweight system with a small form-factor.

In order to function with a single battery, the circuits are designed to operate from supply voltages below 1.5 V, and consume a total of 500 µW. The architecture is shown in Fig. 11.1a. The system includes a low-noise analog front end (AFE), an 8B ADC, and a 100 kb/s wireless transmitter.

11.1 Analog Front End

The AFE provides gain variable from 42 to 78 dB. It uses the closed-loop amplifier described in Chap. 4 as its first stage to achieve low noise. The second stage is a variable-gain amplifier (VGA).

The variable gain amplifier (VGA) is built from a folded-cascode op-amp with a rail-to-rail input-stage and digitally programmable capacitive feedback. Bits in the configuration shift register select one of six feedback capacitors to set the gain. Because the gain settings are logarithmically spaced, it was not possible to use a binary-scaled array to reduce the number of capacitors. While the low-noise amplifier uses pseudo-resistors to set the DC bias point, the VGA uses transconductance ($\mathrm{G_m}$) cells.

The $\mathrm{G_m}$ cells consume a small amount of additional current, and add noise to the VGA output. The additional current is negligible compared to the current required by the op-amps used in the LNA and VGA. The increased noise is the reason why the pseudo-resistors are chosen for the LNA. Because the gain of the LNA attenuates the input-referred noise contribution of the VGA, a noisier feedback element can be tolerated in exchange for improved control over the low-frequency cutoff. Six $\mathrm{G_m}$ cells correponding to the six gain settings are included in the VGA. Because the high-pass frequency corner f_{HP} is $\mathrm{G_m}/2\pi C_F$, the six $\mathrm{G_m}$ cells are needed to keep f_{HP} corner

DOI 10.1007/978-1-4419-6727-5_11, © Springer Science+Business Media, LLC 2011

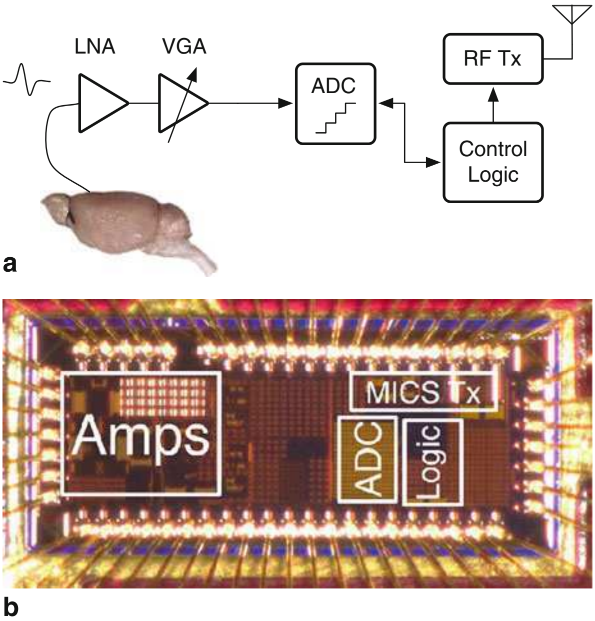


Fig. 11.1 **a** Block diagram of the wireless streaming chip. **b** The neural signal is amplified by the LNA and VGA, digitized, and wirelessly transmitted to a base station

constant over different gain settings. They also allow some tuning of the corner frequency, which can be helpful in balancing the need to reject low-frequency interferers (e.g., 60 Hz wall power, instrumentation noise, local field potentials) with the need to pass the entire band of interest. Alternately, high impedance pseudoresistor feedback can be selected to obtain a high-pass corner below 10 Hz.

The integrated noise from 0.1 Hz to 25.6 kHz is 1.9 μVrms. The power dissipation of the entire analog front-end, including ADC and biasing, is 75 μW, operating from a 1 V supply.

11.2 Conversion and Control

The VGA output is sampled by an 8-bit successive approximation register (SAR) ADC, designed to operate at sample rates from 10 to 100 kS/s. On-chip control logic enables the ADC clock, muxes one of AFE channels to the ADC, and routes the digitized data to the transmitter. The logic block also interleaves alignment data between ADC words to aid in clock/data recovery at the receiver. The alignment header is a "010" string, to ensure that there is at least one transition for every digitized value, simplifying clock recovery. While the header successfully enables reconstruction of the the addition of three bits of overhead for every eight bits of data is inefficient. With the transmitter's data rate of 100 kb/s, the sampling rate is 9.1 kS/s, whereas a sampling rate of 12.5 kS/s would have been possible with no synchronization bits added. Future systems would benefit from using a more efficient coding scheme.

11.3 MICS-band Wireless Transmitter

The transmitter uses binary frequency shift keying (FSK) modulation and is designed to transmit in the Medical Implant Communication Service (MICS) band at 402–405 MHz or the ISM band at 433 MHz. Frequency modulation is accomplished by pulling a crystal reference oscillator using an on-chip capacitor. The crystal oscillator output is then multiplied 9× using a delay-locked loop (DLL) and edge combiner. The edge combiner also drives the antenna through an off-chip impedance matching network, obviating the need for a dedicated power amplifier. The transmitter is described in more detail in [2].

11.4 Results

The recording system was implemented in a $2.5 \times 1\,\text{mm}^2$ die, shown in Fig. 11.1b, using a 0.13 μm CMOS process. The only necessary external components are two quartz crystals for RF carrier generation and system timing, respectively, one inductor for impedance matching, and 6 capacitors for impedance matching, DLL loop filtering, and system clock generation.

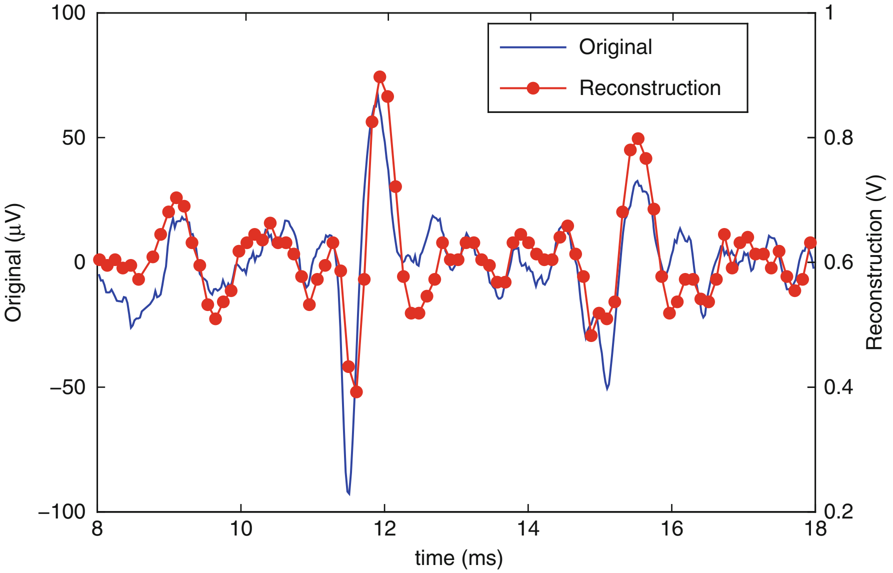


Fig. 11.2 Wirelessly transmitted data. A 160 μV_{PP} artificial neural signal (thin blue trace) was applied to the AFE input, digitized, and transmitted. The signal was received with a commercial FSK receiver located 15 m away. Clock and data were then recovered in software, and compared to the original signal

To verify system functionality, we applied an artificial neural signal with 160 µV$_{PP}$ amplitude to the LNA input. The signal was amplified using the second highest gain setting, digitized, and transmitted over a 15 m wireless link. The RF signal was received using a commercial FSK receiver board [3] and recorded using an oscilloscope. Clock and data recovery and reconstruction of the sampled signal were performed offline in software. Figure 11.2 shows the original input signal and the signal reconstructed from the transmitted data.

We also tested the analog front end in an in vivo recording experiment to verify the compatibility of the amplifiers with the impedance of the recording electrodes. The signal was amplified by the analog front end at the 2nd highest gain setting (72 dB) and digitized with rack-mounted recording equipment. Figure 11.3a shows a 10-s clip of the recorded waveform. Figure 11.3b shows 85 spikes found by a software spike-sorting algorithm [1].

This system can be used for single-channel recording experiments. It demonstrates the suitability of the low-power low-noise amplifier design for neural recording

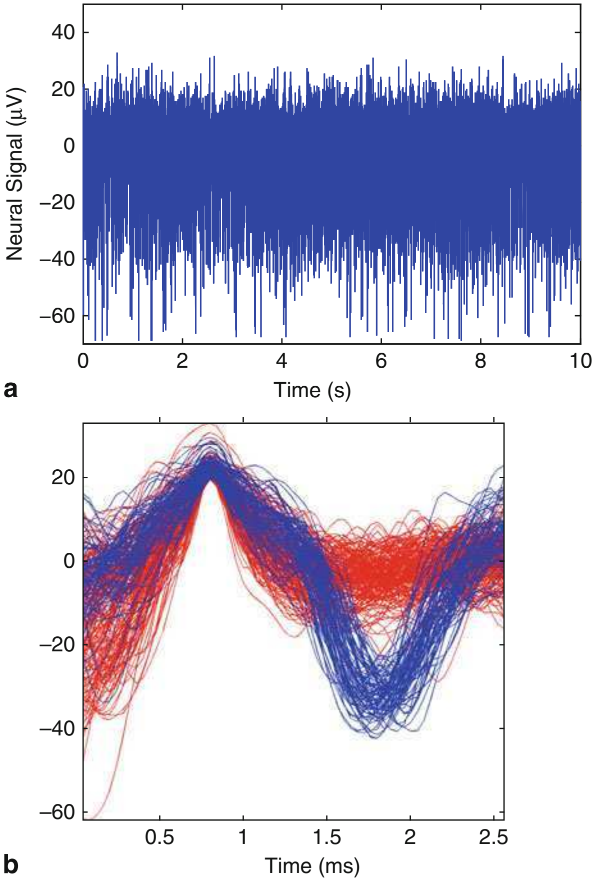


Fig. 11.3 Neural signals recorded in vivo from the motor cortex of a rat. **a** In the 10 s of the signal shown here, many spikes can be seen. **b** Spikes discriminated by a software spike sorting algorithm and overlaid, showing a well-defined characteristic shape

applications. Additionally, it highlights the bottleneck imposed by the wireless transmitter for neural streaming applications. While faster transmitters can improve the situation somewhat, future systems with the ability to record from many channels will require local processing to reduce the required transmitter data rate.

References

[1] Quiroga R, Nadasdy Z, Ben-Shaul Y (2004) Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. Neural Comput 16(8):1661–1687

[2] Rai S, Holleman J, Pandey J, Zhang F, Otis B (2009) A 500 μW neural tag with 2 μVrms AFE and frequency multiplying MICS/ISM FSK transmitter. In: IEEE international solid-state circuits conference, Digest of Technical Papers, pp 212–213

[3] Systems MM (2008) EVB7122 evaluation board. http://www.melexis.com/Hardware_and_Evaluation_Boards/Evaluation_Boards/EVB7122_46.aspx

Chapter 12
Conclusions

Implantable neural interfaces have the potential to revolutionize medicine and neuroscience research. One of the key challenges to realization of this potential is reduction of power consumption. This book has described micro-power circuit implementations for several key building blocks of a neural interface.

A neural recording system of any architecture will require pre-amplifiers with low noise. The amplifiers described in Chaps. 3, 4, and 6 illustrate a variety of approaches to the power-performance tradeoff. Because neural interfaces are typically power-constrained and because the amplifiers can contribute a large fraction of the power dissipation, the power savings provided by these amplifiers can enable improved noise performance, an increased number of channels, or additional functionality. Hopefully the ideas described here will help spur further improvements in amplifier design.

Although the amplifier designs discussed in this book are optimized for power and noise, there is still room for improvement. As most of the work focused on optimizing the power-noise trade-off, we could see the inverse relationship between noise performance and power consumption in the amplifiers. A natural question is whether a given level of noise performance is necessary. The answer, of course, depends on the application. The required noise performance depends on the noise floor of observed signal, the nature of the spiking activity, and the application. In the future designs, the current should be made adjustable depending on the noise requirement of the system to conserve power.

The work presented in this book also represents novel structures for analog computation. Circuits were demonstrated to detect and characterize spikes with power consumption below 1 μW. A floating-gate memory cell utilizing tunneling through thin oxides was demonstrated for the first time. The memory cell was utilized in an analog clustering circuit, which is a key component of a spike sorting system. Similar memory cells could also be used in other neural network applications to allow extremely low-power machine learning.

The results shown in Chap. 7 and 9 demonstrate the feasibility of performing computation in low-power circuits fabricated in modern CMOS processes while highlighting challenges to be addressed in future work. While the analog circuits described here operate with very low power consumption, their functionality cannot be

J. Holleman et al., *Ultra Low-Power Integrated Circuit Design for Wireless Neural Interfaces*, 117
DOI 10.1007/978-1-4419-6727-5_12, © Springer Science+Business Media, LLC 2011

easily modified after chip fabrication. The design of analog computational structures is also more difficult and risky than digital design.

Architectures to add flexibility to the computation performed will make analog computation more competitive with digital solutions. Work in this area has begun [1, 2] and with continued work we can hope to see powerful, flexible, low-power computational engines in the near future.

Improved design methodology can help mitigate the risk and reduce design time for analog computational systems, much as they already have for digital designs. Specifically, the integration of tools to combine of behavioral modeling and transistor-level circuit simulation will make it easier to predict system-level ramifications of circuit-level non-idealities. Improved prediction of device variation and variation-tolerant circuit techniques will also enable improved performance from extremely low-power circuits.

The final two chapters represent initial efforts at integrating some of the ideas presented here into functional neural interfaces. The NeuralWISP project demonstrates how analog signal processing, specifically spike detection, can reduce the computational burden on a digital processor and reduce overall system power. The streaming system demonstrates the practical feasibility of using a complementary input stage to improve noise/power performance in neural recording amplifiers. It also provides an example of how circuit techniques in both the analog front end and communications link can enable a wireless recording system to operate with very low power dissipation.

Although neural recording is the application for which the circuits described here were developed, the techniques should be applicable to the power-constrained signal processing problems encountered in other sensor applications as well. As the need for ultra-low-power operation becomes more widely acknowledged and efficient systems-on-chip become available for sensing tasks, we can expect to see sensing electronics utilized in an ever-growing number of disciplines.

A great deal of work remains to make autonomous neural interfaces practical, and it will be necessary to explore every avenue to find strategies that can provide the needed functionality with the available power and an acceptable form factor. It is our hope that the work described here can contribute to the realization of that potential.

References

[1] Bridges S, Figueroa M, Hsu D, Diorio C (2005) A reconfigurable VLSI learning array. In: Proceedings of conference on the 31st European solid-state circuits conference (ESSCIRC 2005). Grenoble, France, September 2005, pp 117–120

[2] Hall T, Twigg C, Gray J, Hasler P, Anderson D (2005) Large-scale fieldprogrammable analog arrays for analog signal processing. IEEE Trans Circuits Syst-I: Regul Pap 52(11):2298–2307

Index

J. Holleman et al., *Ultra Low-Power Integrated Circuit Design for Wireless Neural Interfaces*, DOI 10.1007/978-1-4419-6727-5, © Springer Science+Business Media, LLC 2011